DEPARTMENT OF ECONOMICS

Working Paper

SPECIFYING NONLINEAR ECONOMETRIC MODELS BY FLEXIBLE REGRESSION MODELS AND RELATIVE FORECAST PERFORMANCE

Christian M. Dahl
Svend Hylleberg

Working Paper No. 1999-4
Centre for Non-linear Modelling in Economics

ISSN 1396-2426

UNIVERSITY OF AARHUS • DENMARK

CENTRE FOR NON-LINEAR MODELLING IN ECONOMICS
DEPARTMENT OF ECONOMICS - UNIVERSITY OF AARHUS - DK - 8000 AARHUS C - DENMARK
☎ +45 89 42 11 33 - TELEFAX +45 86 13 63 34

WORKING PAPER

SPECIFYING NONLINEAR ECONOMETRIC MODELS BY FLEXIBLE REGRESSION MODELS AND RELATIVE FORECAST PERFORMANCE

Christian M. Dahl
Svend Hylleberg

Working Paper No. 1999-4

DEPARTMENT OF ECONOMICS
SCHOOL OF ECONOMICS AND MANAGEMENT - UNIVERSITY OF AARHUS - BUILDING 350
8000 AARHUS C - DENMARK ☎ +45 89 42 11 33 - TELEFAX +45 86 13 63 34

Specifying Nonlinear Econometric Models by Flexible Regression Models and Relative Forecast Performance

Christian M. Dahl and Svend Hylleberg
CNLME, Department of Economics, University of Aarhus
E-mail: Cdahl@econ.au.dk and Shylleberg@econ.au.dk

March 23, 1999

Abstract

The paper considers the task of selecting a flexible nonlinear model which can be used as a baseline model. The baseline model may be used as a testing ground for more structural models which are congruent with economic theory. From the limited empirical evidence obtained here it is tentatively suggested to find a baseline nonlinear flexible form for a univariate time series by following the procedure: 1. Recursively, based on h extra periods at a time specify and estimate a linear form by use of model selection criteria like Cross Validation and/or BIC. 2. After a preliminary test for linearity, recursively, specify and estimate flexible regression models like the FNL suggested by Hamilton (1999) and the Projection Pursuit model suggested by Aldrin, Boelviken and Schweder (1993) for cases of moderate nonlinearities. Use the Cross Validation and the BIC criteria. 3. Based on the remaining part of the data set select the best nonlinear flexible form by use of forecast criteria measuring the absolute forecast performance and the directional forecast performance in h-steps ahead predictions, and compare the best flexible form to the linear specification by use of the Diebold Mariano tests, see Diebold and Mariano (1995) and the forecast encompassing tests suggested by Harvey, Lebourne, and Newbold (1998). The results indicate that the FNL method and the Projection Pursuit Model are the preferable models to apply and that the CV and BIC are the best selction criteria, while the forecast encompassing tests properly modified as suggested by Harvey et. al.(1998) possess better power properties than the Diebold-Mariano test.

JEL Classification: C10, C45, C50

Keywords: Nonlinear model specification

Acknowledgement 1 *An earlier version of the paper was presented as an invited paper at the EC^2 meeting in Stockholm 17.-19.12 1998. Comments and suggestions from the participants and seminar participants at the University of Aarhus, the University of Venezia Ca'Foscari, and the European University Institute in Florens are gratefully acknowledged.*

Contents

1 Introduction **2**

2 Flexible Regression Models **5**
2.1 Hamilton's Flexible Regression Model. 5
2.2 The Neural Network Regression Model 8
2.3 The Projection Pursuit Regression Model 9

3 The choice of the baseline flexible regression model **14**
3.1 Tests for linearity 17
3.1.1 Hamilton's Lagrange Multiplier test 17
3.1.2 The Neural Network test 18
3.1.3 The RESET and Tsay tests 18
3.1.4 White's Dynamic Information Matrix test 19
3.1.5 Application 1a. The change in the US Unemployment rate 19
3.1.6 Application 2a. The growth rate of US industrial production 20
3.2 Model selection and estimation 20
3.2.1 Hamilton's Flexible Regression Model. 21
3.2.2 The Neural Network Regression Model. 21
3.2.3 The Projection Pursuit Regression Model. 21
3.2.4 h steps ahead forecasts 22
3.2.5 Measures of the absolute forecast performance 23
3.2.6 Measures of directional forecast performance 24
3.2.7 Application 1b. The change in the US Unemployment rate 25
3.2.8 Application 2b. The growth rate of US industrial production 26
3.3 Comparing real time forecast abilities 27
3.3.1 The Diebold-Mariano test 27
3.3.2 The forecast encompassing tests. 28
3.3.3 Application 1c. The change in the US unemployment rate 30
3.3.4 Application 2c.The growth rate of US industrial production 30

4 Conclusions **31**

5 Tables **35**

1 Introduction

Due to thresholds, capacity constraints, rationing, institutional restrictions like tax brackets, and asymmetries of different kinds, nonlinear relations are an integral part of many economic theories. Nonetheless, most empirical econometric models are basically linear. Several explanations for this state of the art can be given in addition to the obvious one of familiarity and convenience. One of the possible explanations is based on the procedure often applied when deciding on the application of a nonlinear or linear specification. In most cases a linear model is specified at the outset and a nonlinear specification is only considered if some test for nonlinearity indicates that the linear specification may be in doubt. Unfortunately, very little information as to what kind of nonlinear model should be applied, can be extracted from most or all of the existing tests for nonlinearity implying that the actual choice of nonlinear model is rather arbitrary and tailored to the data in the sample applied. In fact, it often turns out that the out of sample forecast performance of the rejected linear model is far better than the out of sample forecast performance of the nonlinear alternative. As out of sample forecast performance is one of the preferred means to guard against the inherent danger of overparameterized nonlinear models, such evidence typically implies that the nonlinear specification of the model is in doubt. Hence, the linear model often ends up as the preferred specification. On the other hand, even if the nonlinear specification adopted should have better forecast performance than the linear specification, there may exist another nonlinear specifications, which in a better and more parsimonious way represents the theoretical and empirical information available. If this is the case, how do we avoid ending up with a nonlinear specifications which at best has little credibility as it has only been chosen because it performs better than a specific linear model, and maybe is at odds with a more general data coherent nonlinear model?

In accordance with the general to specific modelling strategy, one might argue that the best way to proceed is to start by applying a specification flexible enough to contain the linear and a wide range of nonlinear specifications as special cases. The flexible nonlinear specification can then play the role of a base line model encompassing the class of models within which more specific and interpretable models must be found. The more specific models must then be tested against the general model and not rejected in order to be applied in the subsequent analysis. Two major problems exist. Firstly, how do we determine the base line flexible model in a feasible and cost effective way, and secondly, how do we avoid that the base line model is an overparameterized model, overfitting the sample applied? The answer could be to apply one of the flexible nonlinear regression models available specified in a computer intensive but labour saving way and base the choice of base line model on the relative forecast accuracy of the model in an out of sample context. A related approach to the one suggested here is advocated by Swanson and White (1995, 1997a, 1997b) and Stock and Watson (1998). The approach is applied in finding nonlinear components in US macroeconomic series. Both Swanson and White and Stock and Watson select the preferred model based on the *BIC* criterion, and while Swanson and White applies a neural network nonlinear model Stock and Watson also apply exponential smoothing and smooth transition autoregressions. In both cases the results are quite favourable to the linear model, but we will argue below that this result may be due to the choice of model selection criterion and to the class of nonlinear models used in the comparisons. The baseline model may be chosen among several flexible nonlinear regression models available in the literature .

In this paper we consider flexible nonlinear regression models such as Hamilton's Flexible Regression Model (FNL), see Hamilton (1999) , the Neural Network Regression Model (ANN), see White (1992) , and two versions of the Projection Pursuit Regression Model (PPR). The first version of the PPR model is based on the algorithm suggested by Friedman and Stuelze (1981) (PPR1) and the second is suggested by Aldrin, Boelviken, and Schweder (1993) to be applied in cases of moderate non-linearities (PPR2). The FNL and ANN are parametric models while the PPR models are nonparametric.

Although the applications of the flexible models in the present context are restricted to univariate models the alternative flexible models are chosen among the class of models which can easily be generalized to the multivariate case. The model selection within each of the four classes of models is made by a forward stepwise procedure, where a simultaneous estimation method is applied at each step, and where the model selection criteria applied are the AIC criterion, the BIC criterion or Cross Validation (CV). The results of the applications indicate that the CV criterion and in some cases the BIC criterion should be preferred as they lead to the best forecasting models. The predictive precision of the flexible regression models is evaluated by use of absolute forecast performance measures such as for instance the Mean Square Error and directional forecast measures such as the degree of diagonal concentration. In order to compare the four different models we also apply the Diebold and Mariano test, (DM), see Diebold and Mariano (1995) for relative predictive accuracy and the forecast encompassing test suggested by Harvey, Leybourne, and Newbold (1998), in addition to a test based on Spearman's rank correlation coefficient. In the applications performed below only the forecast encompassing tests seem to have the power to separate nonlinear models from linear models.

Tentatively, the results suggest that the small sample power properties of the forecast encompassing tests are better than the small sample properties of the Diebold-Mariano test, at least when the modelling procedure starts from the linear specification and adds the nonlinear parts. In addition, the size properties of the encompassing tests are quite acceptable, as shown by Harvey et. al. (1998). The procedures are applied to the growth rates in US industrial production and US unemployment in an attempt to make an additional contribution to the ongoing discussion of the possible existence of asymmetries and nonlinearities in the US business cycle[1]. For both series the results indicate that the forecast accuracy of the flexible regression models is in general better than the forecast accuracy of the linear models. Among the non-linear models Hamilton's flexible regression model and a project pursuit model applicable in case of moderate non-linearities seem preferable as the choice for the baseline model. The outline of the presentation is that the flexible regression models are presented in Section 2, while Section 3 contains a presentation of the procedure suggested when choosing the baseline model. The first subsection contains the presentation and application of the tests for linearity applied, while the second subsection contains the discussion of model selection and estimation procedures. The last subsection describes the real time forecast comparison between the best linear and nonlinear models. Section 4 contains the conclusions.

[1] The data may be downloaded from
http://www.econ.au.dk/vip_htm/shylleberg/webpage/shpage.html

2 Flexible Regression Models

Four Flexible Regression Models will be considered. Three of these, the Neural Network Regression Model, see White (1992), and the Projection Pursuit Regression Model , see Friedman and Stueltze (1981), Huber (1985) and Härdle (1990) and the Projection Pursuit Regression Model for moderate nonlinearities, see Aldrin, Boelviken and Schweder (1993) are already well known in the literature although applications in the field of dynamic time series analysis in other areas than financial markets are limited. The fourth approach - denoted Hamilton's Flexible Regression Model - is novel and due to Hamilton (1999). While the Neural Network model and the the Projection Pursuit models specify the nonlinear components as part of the mean function the model suggested by Hamilton introduce the nonlinear components as part of the covariance matrix of the disturbance term.

2.1 Hamilton's Flexible Regression Model.

The basic idea underlying the flexible regression model approach suggested by Hamilton (1999) is to view not only the endogenous variable as a realization of a stochastic process but also to consider the functional form of the conditional mean function itself as the outcome of a random process. Consider the model[2]

$$y_t = \mu_{fnl}(x_t, \delta) + \epsilon_t \tag{1}$$

where ϵ_t is a sequence of $NI(0, \sigma^2)$ error terms and $\mu_{fnl}(x_t, \delta)$ is a function of a $k \times 1$ vector x_t, which may include lagged dependent variables. Let us represent the mean of the conditional distribution i.e. $\mu_{fnl}(x_t, \delta)$ as having a linear part and a stochastic nonlinear part i.e. as [3]

$$\mu_{fnl}(x_t, \delta) = x_t^{'}\beta + \lambda m(g \odot x_t) \tag{2}$$

where for any choice of z, $m(z)$ is a realization from a random field with the asymptotic distribution

$$\begin{aligned} m(z) &\sim N(0,1) && (3) \\ E(m(z)'m(w)) &= H_k(h) && (4) \end{aligned}$$

and where h is defined as $h \equiv \frac{1}{2}[(z-w)'(z-w)]^{\frac{1}{2}}$. The realization of $m(.)$ is viewed as being predetermined with respect to $\{x_1, .., x_T, \epsilon_1, .., \epsilon_T\}$ and $m(.)$ is therefore considered to be independent of $\{x_1, .., x_T, \epsilon_1, .., \epsilon_T\}$. The covariance matrix $H_k(h)$ is defined by

$$H_k(h) = \begin{cases} G_{k-1}(h,1)/G_{k-1}(0,1) & \text{if } h \le 1 \\ 0 & \text{if } h > 1 \end{cases} \tag{5}$$

[2] For the sake of convenience it is assumed in the following that all variables are demeaned.

[3] Here g is a $k \times 1$ vector of parameters and $\odot$ denotes element-by-element multiplication i.e. g$\odot x_t$ is the Hadamard product. β is a kx1 vector of coefficients.

where $G_k(h,r)$, $0 < h \leq r$ is[4]

$$G_k(h,r) = \int_h^r (r^2 - z^2)^{\frac{k}{2}} dz \tag{6}$$

Closed form expressions for $H_k(h)$ for $k = \{1,..,5\}$ are provided by Hamilton (1999) who also gives a general description of the statistical properties of the random field[5].

Since it is not possible to directly observe $m(z)$ - for any choice of z - we cannot observe the functional form of $\mu_{fnl}(x_t, \delta)$. Hence inference about the unknown parameters of the model summarized by $\delta = \{\beta, \lambda, g, \sigma\}$ must be based on observing the realizations of y_t and x_t only. For that purpose rewrite model (1) as

$$y = X\beta + u \tag{7}$$

where y is a $T \times 1$ vector with t'th element equal to y_t , X a $T \times k$ matrix with t'th row equal to $x_t^{'}$ and u a $T \times 1$ vector with t'th element equal to $\lambda m(g \odot x_t) + \epsilon_t$. Conditional on an initial set of parameters λ, g, and by defining $\zeta \equiv \frac{\lambda}{\sigma}$ and $W(X; g, \zeta) = \zeta^2 H + I_T$ we may obtain the GLS estimate of the parameters of the linear part of the model consisting of β and σ as

$$\begin{aligned} \widehat{\beta}_T(g,\zeta) &= [X'W^{-1}(X;g,\zeta)X]^{-1}X'W^{-1}(X;g,\zeta)y & (8) \\ \widehat{\sigma}_T^2(g,\zeta) &= \frac{1}{T}[y - X\widehat{\beta}_T(g,\zeta)]'W^{-1}(X;g,\zeta)[y - X\widehat{\beta}_T(g,\zeta)] & (9) \end{aligned}$$

where I_T is the identity matrix of dimension $(T \times T)$ and the $\{t,s\}$ entry of the matrix H - denoted $H(t,s)$ - is equal to

$$\begin{aligned} H(t,s) &= \begin{cases} H_k(h_{ts}) & \text{if } h_{ts} \leq 1 \\ 0 & \text{if } h_{ts} > 1 \end{cases} & (10) \\ h_{ts} &= \frac{1}{2}[(\widetilde{x}_t - \widetilde{x}_s)'(\widetilde{x}_t - \widetilde{x}_s)]^{\frac{1}{2}} \\ \widetilde{x}_t &= g \odot x_t \end{aligned}$$

Based on the ideas of Wecker and Ansley (1983) , Hamilton (1999) shows, that the concentrated log likelihood function can be written as

$$\mathcal{L}(y, X; g, \zeta) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln \widehat{\sigma}_T^2(g,\zeta) - \frac{1}{2}\ln|W(X;g,\zeta)| - \frac{T}{2} \tag{11}$$

[4] Notice $G_0(h,r) = h - r$, and $G_k(h,r)$ can then be computed recursively by

$$\begin{aligned} G_k(h,r) &= -\frac{h}{1+k}(r^2 - h^2)^{k/2} \\ &\quad + \frac{kr^2}{1+k}G_{k-2}(h,r) \end{aligned}$$

for $k = 2, 3, ...$

[5] The correlation between $m(z_t)$ and $m(w_s)$ is given by the volume of the intersection of a k dimensional unit spheroid centered at z_t and a k dimensional unit spheroid centered at w_s relative to the volume of a k dimensional unit spheroid. Hence, the correlation between $m(z_t)$ and $m(w_s)$ is zero if the Euclidean distance between z_t and w_s is ≥ 2.

The value of the concentrated likelihood function is found using the values of $\widehat{\beta}_T(g,\zeta)$ and $\widehat{\sigma}_T^2(g,\zeta)$ found from equation (8) and (9). A new set of values for g and ζ is selected and a new value of the concentrated likelihood function (11) is found again using the new values of $\widehat{\beta}_T(g,\zeta)$ and $\widehat{\sigma}_T^2(g,\zeta)$ found from equation (8) and (9). Once the estimates of (g,ζ) maximizing equation (11) have been obtained, a new estimate of $\widehat{\beta}_T$ and $\widehat{\sigma}_T^2$ is given from (8) and (9).

From the specification in (7) it is clear that the nonlinearities are introduced into the Hamilton model through the specification of the error covariance matrix.

The estimator of the conditional mean function $\mu_{fnl}(x_t,\delta)$ is given by the t'th row of

$$X\widehat{\beta}_T + \widehat{P}_0(\widehat{P}_0 + \widehat{\sigma}_T^2 I_T)^{-1}[y - X\widehat{\beta}_T] \tag{12}$$

where the $\{t,s\}$ entry of the matrix P_0 - denoted $\mathrm{P}_0(t,s)$ - is equal to

$$\begin{aligned} P_0(t,s) &= \begin{cases} \lambda^2 H_k(h_{ts}) & h_{ts} \leq 1 \\ 0 & h_{ts} > 1 \end{cases} && (13) \\ h_{ts} &= \frac{1}{2}[(\widetilde{x}_t - \widetilde{x}_s)'(\widetilde{x}_t - \widetilde{x}_s)]^{\frac{1}{2}} && (14) \\ \widetilde{x}_t &= g \odot x_t && (15) \end{aligned}$$

as shown by Hamilton (1999).The estimator of the conditional mean function will be consistent for $\mu(.)$ belonging to a very broad class of deterministic nonlinear functions. This result will also apply in the case of $\mu(.)$ being linear. Since we are going to evaluate the forecast accuracy of the model out of sample and equation (12) only works for cases where the conditional mean function is evaluated at points observed in the sample, a modification must be made. To be more specific, in order to obtain the maximum likelihood estimate of μ_{fnl} and δ, denoted $\widehat{\mu}_{fnl}$ and $\widehat{\delta}$ respectively. we seek to calculate $\widehat{\mu}_{fnl}^*(x^*,\widehat{\delta})$,where $x^* = \{x_1^*, x_2^*, .., x_k^*\}$ does not belong to the sample.If we let $P_0^*(t)$ denote the covariance between $\mu_{fnl}(x_t,\delta)$ and $\widehat{\mu}_{fnl}^*(x_t,\widehat{\delta})$, we can obtain an estimate of $\mu_{fnl}^*(x^*,\widehat{\delta})$ as

$$\widehat{\mu}_{fnl}^*(x^*,\widehat{\delta}) = X\widehat{\beta}_T + \widehat{P}_0^{*\prime}(\widehat{P}_0 + \widehat{\sigma}_T^2 I_T)^{-1}[y - X\widehat{\beta}_T] \tag{16}$$

where

$$\begin{aligned} P_0^* &= \{P_0^*(t), t = 1,2,...T\} && (17) \\ P_0^*(t) &= \begin{cases} \lambda^2 H_k(h_t^*) & h_t^* \leq 1 \\ 0 & h_t^* > 1 \end{cases} && (18) \\ h_t^* &= \frac{1}{2}[(\widetilde{x}_t - \widetilde{x}^*)'(\widetilde{x}_t - \widetilde{x}^*)]^{\frac{1}{2}} && (19) \\ \widetilde{x}_t &= g \odot x_t && (20) \\ \widetilde{x}^* &= g \odot x^* && (21) \end{aligned}$$

for $t = 1,..,T$. In equation (16) $\widehat{P}_0$ and $\widehat{P}_0^*$ denote P_0 and P_0^* evaluated at the maximum likelihood estimates of λ and g.

2.2 The Neural Network Regression Model

Neural networks models are nonlinear models that can be specified to fit past and future values of a time series hereby extracting hidden structures and relationships governing the data. In a traditional statistical context neural networks can be considered a nonlinear, non-parametric inference technique that in the unconstrained form is data driven and model free. A priori the relationships between input variables and output variables are unconstrained and no predetermined parameters are required to specify the model. Let us consider the single hidden layer feedforward[6] network in which the output y_t given inputs x_t is determined as

$$y_t = \mu_{ann}(x_t, \kappa) + \epsilon_t \tag{22}$$

where

$$\begin{aligned} \mu_{ann}(x_t, \kappa) &= x_t'\beta + \sum_{j=1}^{q} \theta_j \psi_j(x_t'\gamma_j) & (23) \\ \kappa &= \{\beta, \theta_1, \theta_2, ...\theta_q, \gamma_1, .., \gamma_q\} & (24) \end{aligned}$$

In this type of neural network k input units send the signals x_{it} to so-called "hidden" units across weighted connections γ_{ij} for $i = 1, .., k$ and $j = 1, .., q$. There are in total q hidden units each observing the weighted sum of the k input signals, that is, hidden unit j observes $x_t'\gamma_j$ where $x_t = \{x_{1t}, .., x_{kt}\}$ and $\gamma_j = \{\gamma_{1j}, .., \gamma_{kj}\}$. The hidden unit j then outputs a signal $\psi_j(x_t'\gamma_j)$ where ψ_j denotes the "activation" or "squashing" function commonly assumed to be bounded and monotonically increasing. Following White (1989) we take the activation function to be a logistic function and to be identical for all hidden units, i.e. $\psi_j(x_t'\gamma_j) = \psi(x_t'\gamma_j) = (1 + \exp(-x_t'\gamma_j))^{-1}$ for $j = 1, .., q$. Furthermore, we augment the single hidden layer network by direct links from the input units to a single output with weights $\beta = \{\beta_1, .., \beta_k\}$ implying that the neural network model will have a linear component and assume that the output also contains a white noise term $\epsilon_t \sim nid(0, \sigma^2)$. Finally, we let $\theta = \{\theta_1, .., \theta_q\}$ denote the hidden-units-to-output weights. The parameters of the model and hence the estimate of the conditional mean function is obtained by applying nonlinear least squares, NLS, i.e. by solving

$$\min_{\kappa} E(y - \mu_{ann}(x_t, \kappa))^2 \tag{25}$$

The NLS procedure may converge to a local rather than a global optimum and therefore proper starting values are known to be of a very important matter. For every single specification under consideration we therefore worked with 5 different sets of parameter vectors of starting values and iterated from these until convergence. The iterated parameter vector corresponding to the smallest

[6] The model is denoted feedforward because signals flow from input to output and not vice versa.

value of the objective function given by equation (25) was chosen. As shown by White (1992) the single hidden layer feedforward neural network models possess the universal approximation property that it can approximate any nonlinear function to an arbitrary degree of accuracy with a suitable number of hidden units. Of course this tells us nothing about the performance of such techniques in practice, and for a given set of data it is possible for one technique to dominate another in terms of accuracy etc.

2.3 The Projection Pursuit Regression Model

One very closely related approach to the parametric Neural Network Regression Model is the semiparametric Projection Pursuit Regression Model proposed by Friedman and Stueltze (1981) and Huber (1985), see also Härdle (1990). For a single output variable y_t and a variable input vector given by x_t the Projection Pursuit Regression Model can be written in the form

$$y_t = \mu_{ppr}(x_t, \varrho) + \epsilon_t \tag{26}$$

where

$$\begin{aligned} \mu_{ppr}(x_t, \varrho) &= x_t'\beta + \sum_{j=1}^{v} \omega_j \varphi_j(x_t'\Phi_j) & (27) \\ \varrho &= \{\beta, \omega_1, ...\omega_v, \Phi_1, .., \Phi_v\} & (28) \end{aligned}$$

The parameters Φ_j define the projection of the input vector x_t onto a set of planes labelled by $j = 1, .., v$. These projections are transformed by the nonlinear activation functions denoted $\varphi_j(.)$ and these in turn are linearly combined with weight ω_j and added to the linear part, $x_t'\beta$ to form the output variable y_t.

The first algorithm for obtaining an estimate of ϱ is the original algorithm suggested by Friedman and Stueltze (1981), but with a few differences. Firstly, the algorithm is augmented such that the estimation of the weights ω_j for $j = 1, .., v$ can be obtained using simple ordinary least squares. Secondly, least squares techniques is used for concentrating out the linear part of y. Thirdly, cubic splines with automatic data dependent determination of the smoothing parameter is applied in the estimation of the empirical activation functions φ_j, $j = 1, .., \nu$. Finally, AIC, BIC and CV in turn are used as stopping rules with respect to the choice of the appropriate number of activation functions given by ν, but also as determinants of the number of regressors, k_j for $j = 1, .., \nu$. included in every activation function.

Let y and X be as defined in equation (7). The set of regressors included in the various activation functions is allowed to differ. Let X^j denote the matrix of the k_j regressors included in activation function φ_j. How to choose the proper X^j is of course one of our main concerns and it will be discussed in details later. Furthermore, we define $z^{v+1} = X^{v+1}\Phi_{v+1}$.The algorithm denoted PPR1 can then be described as follows

1. Condition on the regressors by computing the residuals $\widehat{r}^v$ for $v = 0$ and provide initial starting values for the parameters Φ_v and ω_v. In particular

   $$\begin{align} \widehat{r}^v &= y - X^v(X^{v\prime}X^v)^{-1}(X^{v\prime}y) \tag{29} \\ \Phi_v &\sim U(-1, 1) \tag{30} \\ \omega_v &\sim U(-1, 1) \tag{31} \end{align}$$

   where U is the uniform distribution. In order to explore the nature of the local optima repeated runs are important and necessary when estimating projection pursuit models. As for the ANN model we considered 5 different sets of starting values every time a new hidden unit was introduced. These were all drawn from a uniform distribution defined on the unit interval.

2. Find the projection vector $\Phi_{v+1} \in \Re^{k_{v+1}}(||\Phi_{v+1}|| = 1)$ that maximizes the goodness of fit measure $R^2_{v+1}(\Phi_{v+1})$ defined as

   $$R^2_{v+1}(\Phi_{v+1}) = 1 - (\widehat{r}^{v\prime}\widehat{r}^v)^{-1}(\widehat{r}^v - \widehat{\omega}_{v+1}\widehat{\varphi}_{v+1}(z^{v+1}))'(\widehat{r}^v - \widehat{\omega}_{v+1}\widehat{\varphi}_{v+1}(z^{v+1})) \tag{32}$$

   where the estimated empirical activation function $\widehat{\varphi}_{v+1}(z^{v+1})$ is determined by the cubic spline approach

   $$\begin{align} \widehat{\varphi}_{v+1}(z^{v+1}) &= \min_{w(z^{v+1})} S_{\lambda_s}(w) \tag{33} \\ S_{\lambda_s}(w) &= (y - w(z^{v+1}))'(y - w(z^{v+1})) + \lambda_s \int (w''(z^{v+1}))^2 dz^{v+1} \tag{34} \end{align}$$

   with the weight $\widehat{\omega}_{v+1}$ obtained from a linear regression of $\widehat{r}^v$ on $\widehat{\varphi}_{v+1}(z^{v+1})$ as

   $$\widehat{\omega}_{v+1} = (\widehat{\varphi}_{v+1}(z^{v+1})'\widehat{\varphi}_{v+1}(z^{v+1}))^{-1}(\widehat{\varphi}_{v+1}(z^{v+1})'\widehat{r}^v) \tag{35}$$

   The basic idea underlying the cubic spline smoothing approach is to produce a flexible curve that provides a good fit of the data, but without possessing too much local variation. In the current context this approach translates into choosing the activation function φ_j as the functional form $w(z)$ that minimizes the Euclidean distance to y but with a penalty for local variation. Without any penalty included in choosing $w(z)$ we would obtain a perfect fit to the data. In the cubic spline smoothing approach the class of curves that provides a perfect interpolation is avoided by introducing a so-called *roughness penalty* for rapid local variation. There are a number of ways to quantify local variation but the cubic spline approach uses the second order derivative, and the *roughness penalty* is given by

   $$\int w''(z)^2 dz \tag{36}$$

which corresponds to the second term in the expression for S_{λ_s}. Hence, the empirical activation function φ_j applied is the function $w(z)$ that minimizes a weighted sum of residual errors and local variation where the weight associated with the local variation is given by the smoothing parameter λ_s.The smoothing parameter λ_s in the univariate cubic spline regression function is determined conditional on a training data set (to be defined later) and according to the generalized cross validation principle coined by Wahba (1977).

3. If R^2_{v+1} is sufficiently small, stop the algorithm, otherwise go to step 4

4. Construct a new set of residuals

$$\widehat{r}^{v+1} = \widehat{r}^{v} - \widehat{\varphi}_{v+1}(z^{v+1}) \tag{37}$$

and add an additional activation function. Furthermore update $v = v + 1$ and go back through step 2 and step 3.

One important difference between the neural network model described above and the projection pursuit model of this section is that each hidden unit in the projection pursuit regression is allowed a different activation function and in particular that these functions are not prescribed in advance, but are determined from the data as part of the "training" or estimation procedure. Another difference is that the parameters in the projection pursuit regression are optimized cyclically in groups while those in the neural network are optimized simultaneously. Specifically, estimation in the Projection Pursuit Regression Model takes place for one hidden unit at a time, and for each hidden unit the second-layer weights are optimized first, followed by the activation function and the first layer weights. The process is repeated for each hidden unit in turn until a sufficiently small value of the error function is achieved or until some other stopping criterion is satisfied. Since the output y_t depends linearly on the second-layer parameters, these can be optimized by linear least square techniques. Optimization of the activation functions $\varphi_j(.)$ represents a problem of one-dimensional curve fitting for which a variety of techniques can be used, as for instance the one based on cubic splines or the one based on the Nadaraya-Watson type kernel smoother, see Härdle (1990) for a discussion. Here we have chosen to work with the cubic spline smoother.

Finally we will consider a flexible regression model denoted PPR2 which is very closely related to the PPR model outlined above. However, it is much simpler at the expense of being not quite as flexible as the standard algorithm. Aldrin, Boelviken and Schweder (1993) argue that nonlinear structures in practice often are only moderately deviant from a linear structure. These moderate nonlinear structures include S-shapes and other moderate curvatures, slight jumps in derivatives or local dips or bumps. Upon this argument they are able to implement a very simple and fast algorithm for estimation of the conditional mean function given by equation (20) under the assumption that it is nonlinear but monotone or approximately monotone. Aldrin, Boelviken and Schweder

(1993) provide extensive numerical evidence showing that the PPR2 model outperform the PPR1 model when the nonlinearity is moderate and the signal to noise ratio is small. In economics the use of aggregated data is quite common and as the degree of nonlinearity can diminish with aggregation, see Granger and Teräsvirta (1993) a projection pursuit model like PPR2 is an obvious possibility.

Consider a simple version of the model given by equation (26) - (28) such as

$$\begin{aligned} y &= \varphi(X\Phi) + \epsilon \\ &= \varphi(z) + \epsilon \end{aligned} \tag{38}$$

where y and ϵ are vectors with y_t and ϵ_t as the $t'th$ elements, respectively. And assume without loss of generality that the columns of the matrix X with the $t'th$ row equal to x_t' have zero mean and that the coefficient vector Φ is standardized so that $\Phi'\Sigma_x\Phi = 1$ where $\Sigma_x = E(X'X)$. Furthermore, denote the residual vector as $r = y - X'\Phi^{ols}$ where $\Phi^{ols} = \Sigma_x^{-1}\Sigma_{xy}$ where $\Sigma_{xy} = E(X'y)$. In addition, let the predictor be $z = X\Phi$, and consider the vector $u = (X - z\Phi'\Sigma_x)$ and notice that z and u are uncorrelated as $E(z'u) = E(z'(X - z\Phi'\Sigma_x)) = \Phi'\Sigma_x - \Phi'\Sigma_x\Phi\Phi'\Sigma_x = \Phi'\Sigma_x - \Phi'\Sigma_x = 0$. Premultiplying $y = \varphi(X\Phi) + \epsilon$ by X' then imply

$$\begin{aligned} X'y &= X'\varphi(z) + X'\epsilon \\ &= X'\varphi(z) + u'\varphi(z) - u'\varphi(z) + X'\epsilon \\ &= \Sigma_x\Phi z'\varphi(z) + u'\varphi(z) + X'\epsilon \end{aligned} \tag{39}$$

and by applying the expectation operator we obtain

$$\Sigma_{xy} = \Sigma_x\Phi E(z'\varphi(z)) + E(u'\varphi(z)) \tag{40}$$

Finally, premultiplication by Σ_x^{-1} gives

$$\Phi^{ols} = \eta\Phi + B \tag{41}$$

where

$$\eta = E[z'\varphi(z)] \tag{42}$$

$$B = \Sigma_x^{-1}E[u'\varphi(z)] = \Sigma_x^{-1}E[(X - z\Phi'\Sigma_x)'\varphi(z)] \tag{43}$$

Equation (41) writes the OLS estimate Φ^{ols} as a sum of one term proportional to the true direction vector Φ and another term B. As u and z are uncorrelated, the correlation between u and the transformation $\varphi(z)$ may be expected to be small. Indeed, it can be shown, see Aldrin, Boelviken and Schweder (1993), that if X follows an elliptically contoured distribution like the Gaussian, $E[u'\varphi(z)] = 0$ whereby $B = 0$ and $\Phi^{ols} = \eta\Phi$.Also notice that if $\varphi(z)$ is linear i.e. $\varphi(z) = X\Phi$ then $\eta = E(z'\varphi(z)) = \Phi'\Sigma_x\Phi = 1$ and the ordinary least square estimator will be equal to the true parameter $\Phi^{ols} = \Phi$. Heuristically, $\eta = E[z'\varphi(z)]$ measures the correlation between the linear model $z = X\Phi$ and the nonlinear model $\varphi(X\Phi)$. If this correlation is low then Φ^{ols} must be expected to be far from Φ, but if the correlation is reasonably high, which is true in the case of moderate nonlinearities, Φ^{ols} must be expected to be close to Φ. This suggests to take the linear ordinary least squares estimate $\widehat{\Phi}^{ols}$ for Φ and then obtain $\widehat{\varphi}(\widehat{z})$ by smoothing y or in the general case where we have a linear part as in equation (26) r in this direction.

To sum up the simple algorithm based on the ordinary least square estimator will work as long as 1.) η is not to close to zero implying that only moderate nonlinear structures can be analyzed and 2.) B is of a small magnitude requiring that X should be close to but not necessarily perfectly Gaussian distributed.

The first step in the modified algorithm suggested by Aldrin et al. (1993) is to construct a consistent sample version of $\Phi^{ols}_j\eta_j$ and B_j – denoted $\widehat{\Phi}^{ols}_j\widehat{\eta}_j$ and $\widehat{B}_j$ - for $j = 1, .., v$ in order to get a consistent estimate of the true direction vector Φ_j. In our setup the sample version of equation (41) equals

$$\widehat{\Phi}^{ols}_j = \widehat{\eta}\Phi_j + \widehat{B} + (X^{j\prime}X^j)^{-1}(X^{j\prime}\epsilon) \tag{44}$$

$$\widehat{\eta}_j = T^{-1}\widehat{z}_j'\widehat{\varphi}_j(\widehat{z}_j) \tag{45}$$

$$\widehat{B}_j = T^{-1}(X^{j\prime}X^j)^{-1}(\widehat{\varphi}_j(\widehat{z}_j)'(X^j - (X^{j\prime}X^j)^{-1}\widehat{\Phi}^{ols}_j)\widehat{z}_j) \tag{46}$$

$$\widehat{z}_j = X^j\widehat{\Phi}^{ols}_j \tag{47}$$

If the conditions discussed above are fulfilled to an extent such that $\widehat{\Phi}^{ols}_j$ is already close to the true direction, Aldrin et al. (1993) suggest the following iterative scheme moderately changed in order to make it fit into our general model selection strategy

1. Center the response $\widehat{r}^v$ for $v = 0$. The response is centered by de-meaning y by a linear combination of the regressors X^v. In particular,

$$\widehat{r}^v = y - X^v(X^{v\prime}X^v)^{-1}(X^{v\prime}y) \tag{48}$$

2. Estimate Φ^0_{v+1} by ordinary least squares as

$$\widehat{\Phi}^0_{v+1} = (X^{v+1\prime}X^{v+1})^{-1}(X^{v+1\prime}\widehat{r}^v) \tag{49}$$

   and obtain $\widehat{\varphi}_{v+1}$ by using the cubic spline smoother defined by equation (33) with the smoothing parameter determined by generalized cross validation conditional on a training set to be described later.

3. compute $\widehat{\eta}^j_{v+1}$and $\widehat{B}^j_{v+1}$ according to equation (45) and (46) and update the direction vector according to

$$\widehat{\Phi}^j_{v+1} = \frac{\widehat{\Phi}^{j-1}_{v+1} - \widehat{B}^j_{v+1}}{\widehat{\eta}^j_{v+1}} \tag{50}$$

   and standardize such that $\widehat{\Phi}^j_{v+1}(X^{v+1\prime}X^{v+1})^{-1}\widehat{\Phi}^j_{v+1} = 1$.

4. Find the smoothing function $\widehat{\varphi}^j_{v+1}(X^{v+1}\widehat{\Phi}^j_{v+1})$ and obtain the associated optimal weight as

$$\widehat{\omega}^j_{v+1} = (\widehat{\varphi}^j_{v+1}(X^{v+1}\widehat{\Phi}^j_{v+1})'\widehat{\varphi}^j_{v+1}(X^{v+1}\widehat{\Phi}^j_{v+1}))^{-1}(\widehat{\varphi}^j_{v+1}(X^{v+1}\widehat{\Phi}^j_{v+1})'r^v) \tag{51}$$

5. Repeat step 3 and 4 until the scalar difference given by

$$D = || \frac{\widehat{\Phi}^{j-1}_{v+1}}{\sqrt{\widehat{\Phi}^{j-1}_{v+1}(X^{v+1\prime}X^{v+1})^{-1}\widehat{\Phi}^{j-1}_{v+1}}} - \frac{\widehat{\Phi}^{j}_{v+1}}{\sqrt{\widehat{\Phi}^{j}_{v+1}(X^{v+1\prime}X^{v+1})^{-1}\widehat{\Phi}^{j}_{v+1}}} || \tag{52}$$

becomes sufficiently small.

6. If the criterion of fit is satisfied stop the algorithm, otherwise go to step 7

7. Construct a new set of residuals

$$\widehat{r}^{v+1} = \widehat{r}^v - \widehat{\varphi}_{v+1}(X^{v+1}\widehat{\Phi}^j_{v+1}) \tag{53}$$

and add an additional activation function. Furthermore update $v = v + 1$ and go back through step 2 and step 7.

In terms of representational capability, we can regard the standard Projection Pursuit Regression as a generalization of the multilayer neural network model, in that the activation functions are more flexible than those applied in the neural network context. It is therefore not surprising that Projection Pursuit Regression Models should have the same universal approximation capabilities as Neural Network Regression Models.

3 The choice of the baseline flexible regression model

In order to illustrate and compare the use of flexible regression models and relative forecast performance in specifying nonlinear econometric models two data series have been analyzed. The first series is the quarterly change in the US unemployment rate from 1949.Q3 to 1998.Q2, seasonally adjusted, and the second series is the seasonally adjusted quarterly growth rates of US industrial production from 1947.Q2 to 1998.Q2. Both series are seasonally adjusted by X-11. Univariate models using lagged values of the dependent variables as explanatory variables are formulated and the following procedure applied. Firstly, as a preliminary step, the null hypothesis of the series being linear is tested by use of Hamilton's Lagrange Multiplier test, the Neural Network test, the Tsay test, Whites dynamic information matrix test and the RESET test. The number of lags in the linear specification is determined by each of the three model selection criteria applied i.e. AIC, BIC or the Cross Validation (CV).

The first step is undertaken in order to obtain some preliminary information as to the linearity or nonlinearity of the appropriate model, but in the context of this investigation information about the different tests may be obtained as well.

Secondly, the recursive model selection and estimation procedures for each of the four models i.e. Hamilton's Flexible Regression Model (FNL), the Neural Network Regression Model (ANN) , and two versions of the Projection Pursuit Regression Model (PPR1) and (PPR2) are applied and the best specifications chosen based on the AIC, BIC or CV criterion. The forecast ability in a one period ahead forecasting exercise for each of the three versions of the four models is evaluated using measures of the forecasting ability such as the mean square error (MSE), the mean absolute deviation (MAD), the forecast absolute percentage error ($MAPE$), directional tests based on a contingency table such as the Henriksson-Merton test (HM) and the χ^2 test for independence (χ^2), the confusion rate (CR), and the degree of diagonal concentration (ϕ). In addition we report Theil's U statistic (U) and the Granger- Newbold version of the Mincer-Zarnowitz regression, where the actual value is regressed on the forecast, and the coefficient of determination (R ^{2}) applied as a measure if the regression has an intercept of zero and a slope of 1. The objective in the second step is to find the best model based on the precision of the out of sample forecast. In the context of the analysis performed we may also obtain information about the different evaluation criteria AIC, BIC or CV.

Thirdly, the best forecasting model specification from the four classes of nonlinear models is compared to the linear model by use of the Diebold-Mariano test and the forecasting encompassing test suggested by Harvey et al. Model selection will be based on Akaike's and Schwarts' information criterion denoted AIC and BIC respectively and on cross validation denoted CV, see Akaike (1969), Schwarts (1978) , Stone (1974, 1977) and Wahba and Wold (1975). The information criteria AIC and BIC of Akaike(1969) are found as

$$AIC = \ln(\widehat{\sigma}^2) + \frac{2c}{T} \tag{54}$$

$$BIC = \ln(\widehat{\sigma}^2) + \frac{c\ln(T)}{T} \tag{55}$$

where $\widehat{\sigma}^2$ is the estimated variance of the residual from the regression model, c equals the complexity criterion, i.e. the number of estimated parameters and T denotes the sample size. The selected model is the model where c minimizes the criterion. Basically, both criteria are fit criteria with a penalty for the number of parameters. Since our goal is to find the model having the best performance within sample as well as out of sample a simple approach when comparing different model specifications is to evaluate the error function using data which is independent of the sample applied in the estimation part[7].

Cross Validation may be described in the following way. The models to be compared are estimated by minimizations of the appropriate error function

[7] The data set applied in the estimation is also called the training set.

defined on the training data set. The performances of the models are then compared by evaluating the error function using an independent validation set, and the model having the smallest error with respect to the validation set is selected as the preferred model. This approach is called the Hold Out method. The danger of the procedure is that it can lead to overfitting with respect to the validation data set, but we will be able to evaluate the size of the overfitting because we are going to test the performance of the chosen model on a third independent data set namely the out of sample data set. More schematically, the procedure may be described in the following way. First, divide the data set into $S+1$ distinct segments where each of the first S segments consists of [8] $[T/S]$ data points and the remaining segment consisting of $T - S * [T/S]$ data points

$$w = \{y, x\} = \ \{(y^1, x^1), (y^2, x^2), .., (y^S, x^S), (y^R, x^R)\} \tag{56}$$

and

$$\begin{aligned} w &= \{w_t\}_{t=1}^T & (57) \\ w^s &= \{w_{(s-1)*S+1}, .., w_{s*S}\} \ for \ s = 1, .., S & (58) \\ w^R &= \{w_{[T/S]*S+1}, .., w_T\} & (59) \end{aligned}$$

Furthermore, let us introduce the leave-out operator " $\backslash$ " defining the set $\{y^{\backslash s}, x^{\backslash s}\} = \{y, x\} \backslash \{y^s, x^s\}$ as all of the observed data points except the $[T/S]$ observations in segment s. Secondly, estimate the regression model using the data set $\{y^{\backslash s}, x^{\backslash s}\}$ and obtain an estimate of the conditional mean function denoted $\mu(x^{\backslash s}, \widehat{\varsigma}_s)$. Thirdly, evaluate the mean squared regression error function using the remaining $[T/S]$ data points in segment s, that is calculate

$$CV(s) = [T/S]^{-1} \sum_{i=1}^{[T/S]} (y_i^s - \mu_i(x^s, \widehat{\varsigma}_s))^2 \tag{60}$$

where y_i^s and μ_i denote the i'th element/row of the y^s-vector and the μ-vector, respectively. The process is repeated for each value of $s = 1, .., S$, i.e. for every possible choice of segment omitted from the training process, and finally the test errors are averaged over all S . Such a procedure allows us to use a high proportion of the available data (a fraction $1 - 1/S$) to estimate the regression model, while also making use of all data points in evaluating the cross validation error. The disadvantage of such an approach is that it requires the estimation process to be repeated S times, which implies a relatively long processing time. A typical choice of S is $S = 10$, see Bishop (1995) . The single measure applied in the cross validation model selection procedure can be written as

$$CV = S^{-1} \sum_{s=1}^{S} CV(s) \tag{61}$$

[8] Let $[T/S] = \sup\{t : t \leq T/S\}$ for $t = 1, .., T$ i.e that$[T/S]$ the integer on or just below the value of T/S.

3.1 Tests for linearity

As a first step we suggest testing the null hypothesis of the series being linear by use of the most general specifications test. This collection of test statistics will include Hamilton's test, Hamilton (1999), the Regression Error Specification Test or RESET test, Ramsey (1969), a test , denoted the Tsay test, see Tsay (1986), the Neural Network test, see Lee, White and Granger (1993) and a particular version of White's information matrix test, White (1987,1992). The tests are all recommended because of their relatively good performance with respect to size and power against an unspecified alternative model described in the literature. The number of lags in the linear specification under the null is determined by each of the three model selection criteria applied i.e. AIC, BIC or Cross Validation (CV).

3.1.1 Hamilton's Lagrange Multiplier test

The Lagrange multiplier test suggested by Hamilton considers the null hypothesis $H_0 : \lambda = 0$ in equation (2). However, g is not identified when λ equals zero, but Hamilton (1999) solves this problem by assuming that the i'th element of g, i.e. g_i, is proportional to the standard deviation of the i'th row in x_t. Fixing the nonidentified parameters to the scale of the variables implies that the Lagrange multiplier statistics for neglected nonlinearity becomes

$$HLM = \frac{[\widehat{\epsilon}' H \widehat{\epsilon} - \widetilde{\sigma}^2 tr(MHM)]^2}{\widetilde{\sigma}^4 [2tr\{[MHM - (T-k)^{-1} M tr(MHM)]^2\}]} \tag{62}$$

where

$$\begin{aligned} \widehat{\epsilon} &= My && (63) \\ \widetilde{\sigma}^2 &= (T-k)^{-1} \widehat{\epsilon}' \widehat{\epsilon} && (64) \\ M &= I_T - X(X'X)^{-1}X' && (65) \end{aligned}$$

and the (t, s) element of the $T \times T$ covariance matrix H of $m(g \odot x_t)$ is given by

$$\begin{aligned} H(t,s) &= \begin{cases} H_k(h_{ts}) & h_{ts} \leq 1 \\ 0 & h_{ts} > 1 \end{cases} && (66) \\ h_{ts} &= \frac{1}{2}[k^{-1} \sum_{i=1}^{k} \frac{(x_{i,t} - x_{i,s})^2}{s_i^2}]^{\frac{1}{2}} && (67) \\ s_i^2 &= T^{-1} \sum_{t=1}^{T} (x_{i,t} - T^{-1} \sum_{t=1}^{T} x_{i,t})^2 && (68) \end{aligned}$$

where $H_k(.)$ is defined in equation (5). The Lagrange multiplier statistics HLM is asymptotically $\chi^2(1)$ distributed.

3.1.2 The Neural Network test

When the null hypothesis of linearity is true i.e. $\mathrm{H}_0 : \Pr[E(y_t|X_t) = x_t'\beta^*] = 1$ for some choices of β^* and $X_t = \{x_1', x_2', .., x_t'\}$ the optimal network weights θ_j in equation (23) are zero for $j = 1, .., q$. The neural network test for neglected nonlinearity can therefore be interpreted as testing the hypothesis $H_0 : \theta_1 = \theta_2 = .. = \theta_q = 0$ for particular choices of q and γ_j. As in Lee et al. (1993) we set the number of hidden units, i.e. q, equal to 10 and draw the direction vectors γ_j independently from a uniform distribution on the interval [-2:2]. The test is then carried out by an auxiliary regression where[9] $\widehat{\epsilon}_{(T\times1)} = y_T - X_T(X_T'X_T)^{-1}(X_T'y_T)$ is regressed on $\Psi_{(T\times q)} = \{\psi(X_T\overline{\gamma}_1)_{(T\times1)}, .., \psi(X_T\overline{\gamma}_q)_{(T\times1)}\}'$ where $y_T = \{y_1, y_2, .., y_T\}$. The LM-test statistics is given by

$$NNLM = TR^2 \to \chi^2(q) \tag{69}$$

where R^2, is the coefficient of determination from the auxiliary regression. Because the observed components of Ψ_t typically are highly correlated Lee et al. (1993) recommend using a small number of principal components instead of the q original variables. Using the $q^* < q$ principal components of Ψ_t , denoted Ψ_t^*, not collinear with x_t an equivalent test statistics is given by

$$NNLM^* = TR_{pc}^2 \to \chi^2(q^*) \tag{70}$$

where R_{pc}^2 is the coefficient of determination from a regression of $\widehat{\epsilon}_{(T\times1)}$ on $\Psi^*_{(Txq^*)}$.

3.1.3 The RESET and Tsay tests

Consider the linear model

$$y_t = x_t'\beta + u_t \tag{71}$$

where y_t is the dependent variable and x_t a k vector of regressors [10]. The first step consists of regressing y_t on x_t in order to obtain an estimate of β, say $\widehat{\beta}$, the prediction $f_t = x_t'\widehat{\beta}$, and the residuals $\widehat{u}_t = y_t - f_t$ whereby the sum of squared residuals are $SSR_0 = \sum_{t=1}^T \widehat{u}_t^2$. In the second step, regress $\widehat{u}_t$ on x_t and on the sx1vector M_t , to be defined later, and compute the residuals from this regression $\widehat{v}_t = \widehat{u}_t - x_t'\widehat{\alpha}_1 - M_t'\widehat{\alpha}_2$ and the residual sum of squares $SSR = \sum_{t=1}^T \widehat{v}_t^2$. Finally, in the third step compute the F statistics given by

$$F = \frac{(SSR_0 - SSR)/m}{SSR/(T-k-m)} \sim F(s, T-k-s) \tag{72}$$

Under the linearity hypothesis the F statistics above is approximately F-distributed with s and $T-k-s$ degrees of freedom. The difference between the RESET test and the Tsay test lies in the choice of M_t. The RESET test defines M_t as

[9] Notice, that the regressions discussed here all contain an intercept unless otherwise noted.

[10] Notice, the regressors may be lagged dependent variables

$M_t = \{f_t^2, .., f_t^{s+1}\}$. Because f_t^i, $i = 2, .., s+1$ tends to be highly correlated with x_t and with themselves the test is conducted using the $s^* < s$ largest principal components of $f_t^2, .., f_t^{s+1}$ not perfectly collinear with x_t and therefore not with the linear combination $f_t = x_t'\widehat{\beta}$. Tsay (1986) suggests using $M_t = vech(x_t x_t')$ where the operator *vech* implies that M_t contains the elements on and below the diagonal of the matrix $x_t x_t'$ i.e. the squared explanatory variables and the crossproducts of these.

3.1.4 White's Dynamic Information Matrix test

The information matrix test is developed from the observation that if a model is well specified the information matrix equality[11] holds, while this is not the case in a misspecified model. The version of White's dynamic misspecification test considered in this paper will be based on the covariance of the conditional score functions. For a Gaussian linear model the log likelihood function can be written as

$$\mathcal{L}_t(x_t, \beta, \sigma) = -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2}u_t^2 \tag{73}$$

where $u_t = \sigma^{-1}(y_t - x_t'\beta)$. The conditional score function is then given by

$$s_t(x_t, \beta, \sigma) = \sigma^{-1}(u_t, u_t x_t', u_t^2 - 1)' \tag{74}$$

Evaluating the conditional score at the quasi maximum likelihood estimators of the correctly specified model under H$_0$ gives $\widehat{s}_t = s_t(x_t, \widehat{\beta}, \widehat{\sigma})$. The information matrix test is based on forming the qx1 indicator $\widehat{m}_t = S * vec(\widehat{s}_t \widehat{s}_t')$ where S is a selection matrix. In particular we obtain the test statistics denoted "White3" in Lee et. al.(1993) by the auxiliary regression $\widehat{u}_t = \widehat{\sigma}^{-1}(y_t - x_t'\widehat{\beta})$ on x_t and $\widehat{k}_t$ - where $\widehat{k}_t$ is defined to satisfy $\widehat{m}_t = \widehat{k}_t \widehat{u}_t'$. . The test statistic and its asymptotic distribution is then given by

$$WIM = TR^2 \rightarrow \chi^2(q) \tag{75}$$

where R^2 is the coefficient of determination from the auxiliary regression.

3.1.5 Application 1a. The change in the US Unemployment rate

We begin by considering the series of first differences in US unemployment. In order to obtain the correct size of the test the tests for linearity should be based on the residuals from the best linear model as suggested by Granger and Teräsvirta (1993). In practice, this is done by calculating each of the test statistics based on the best linear model being selected by the three model

[11] In a well specified model we can compute the information matrix as minus the expected value of the Hessian matrix of second order derivatives of the log likelihood function or as the outer product of the score vectors i.e. the vector of first order derivatives of the log likelihood function. The relationship which holds under quite general conditions is called the information matrix equality

selection criteria in turn. Furthermore we are conditioning the test statistics on the whole sample period.

We find that the best linear model based on AIC and CV consists of a constant term and 4 lags whereas the best linear model chosen by the BIC criterion includes a constant term but only 2 lags. The results presented in Table 1 indicate that the null hypothesis of linearity in all cases is rejected at the 5% level except in the case where inference is based on the outcome from the RESET test. However, based on the RESET test, rejection of linearity is still supported at a 10% level. Hence, the applied tests indicate the necessity for a non-linear specification of the univariate model for the change in unemployment.

3.1.6 Application 2a. The growth rate of US industrial production

For the growth rate of US industrial production the three model selection criteria agree on the best univariate linear model which includes a constant and two lags, see Table 2. Based on Hamilton's linearity test and the RESET test it is not possible to reject the null of linearity. This result, however is in strict disagreement with the outcome of the Neural Network test, Tsay's and White's test. It could be due to low power of Hamilton's test and the RESET test against a specific kind of nonlinearity inherited in the industrial production series or it could be due to moderate nonlinearities difficult to detect due to robustness and good approximation properties of the linear model one step ahead. The power properties of Hamilton's test always turn out to be at least as good as the power properties of the neural network test, as shown by Dahl (1999) based on a wide range of nonlinear models.

The results presented in Table 2 indicate that the null hypothesis of linearity in the growth rate of US industrial production is rejected by most of the applied tests at the 5% level but clearly not by Hamilton's test and not by the RESET test unless the level of significance is raised to 15%. Hence, the applied tests disagree on the necessity for a non-linear specification of the univariate model for the growth rates of industrial production, and the evidence could imply that the nonlinearity in the US industrial production is nonexistent or of a moderate nature.

3.2 Model selection and estimation

In the second step the recursive model selection and estimation procedures for each of the four models i.e. Hamilton's Flexible Regression Model (FNL), the Neural Network Regression Model (ANN), and two versions of the Projection Pursuit Regression Model (PPR1) and (PPR2) are applied and the best specifications chosen as described above.

Although modern computers are very efficient the computations involved in the different nonlinear approaches discussed here can be excessive. Furthermore, the procedure applied must be somewhat automatic, and in addition parsimony is an important objective. The procedure applied here in the specification and estimation of linear and Flexible Regression Models is a forward stepwise procedure, where a simultaneous estimation method is applied at each step. The exact procedure applied is somewhat dependent upon which of the flexible approaches we apply and we will therefore provide a detailed description of the procedure applied in the three cases.

3.2.1 Hamilton's Flexible Regression Model.

From equation (1) and (2) it is seen that the model contains a linear and a nonlinear part. The first step consists of performing a forward stepwise linear regression with regressors (lags in the univariate case) added on one at a time until no additional regressor improves upon the model selection criterion applied. The number of regressors in the linear part is then fixed. Next the number of regressors in the nonlinear part, the m(.) function, is to be determined. As in the linear part this is done by including regressors one at a time until the model selection criterion cannot be improved upon. If, after adding the first regressor to the nonlinear part of the model , the model selection criterion is not improved upon this will imply that the preferred model will be linear. If applied recursively to different but consecutive time periods the Flexible Regression Model approach allows for the preferred model to be linear in some periods and nonlinear in others. Furthermore, a key feature of the model selection and estimation procedure is that every time a new regressor is added to the model all the parameters in the linear and nonlinear part are reestimated by maximum likelihood.

3.2.2 The Neural Network Regression Model.

First, the number of regressors in the linear part of the model is determined and fixed in exactly the same way as described above. Secondly, a single hidden unit is added and regressors are selected one by one as part of the first hidden unit until the model selection criterion no longer can be improved. The number of regressors included in the first hidden unit is thereafter fixed and a second hidden unit is added and the process repeated until five hidden units have been tried or the model selection criterion cannot be improved upon by adding additional hidden units. Again all the parameters of the model are reestimated by nonlinear least square every time a new regressor is included.

3.2.3 The Projection Pursuit Regression Model.

The model selection procedure applied in connection with the Projection Pursuit Regression Model is similar to the model selection in the neural network case. However, there is one main difference. Since the parameters of the model are estimated in groups, not all the parameters of the model are reestimated every time a new regressor is included. In order to cut down the computational burden we did not consider backfitting as it is also apparent from the description of the PPR1 and PPR2 algorithms. This rule implies that only the set of parameters in the hidden unit in which the new regressor is added are reestimated. The model selection procedure is as follows. Add a hidden unit - which in this case is an empirically determined univariate function - including a constant term and one regressor. Add regressors to this hidden unit and reestimate the model every time a new regressor is included until the model selection criterion cannot be improved. When the number of regressors in the first hidden unit is determined fix both the number of regressors and the parameters at their estimated values. Add the second hidden unit and repeat the process until five hidden units have been tried or the model selection criterion cannot be improved upon by adding additional hidden units.

3.2.4 *h* steps ahead forecasts

In order to evaluate the forecast ability of the three approaches an out of sample one step ahead forecast $\widehat{y}_{t_1+1}$is generated from each of the three flexible regression models estimated by use of a data window containing a sample from the starting point in period t_0 to period t_1. In the next step a second set of one step ahead forecast $\widehat{y}_{t_1+2}$ is computed using a data window beginning at time t_0 and terminating at time $t_1 + 1$. Continuing this procedure, rolling the data window forward one period every time enables us to simulate a sequence of true out of sample forecasts. The sequence we generate contains n data points . The forecast period must be long enough to include periods where the nonlinear characteristics are present. For instance, if asymmetric dynamics over the phases of the business cycle is the expected cause of the nonlinearities, recessions as well as expansionary economic phases must exist in the out of sample forecast period.

In the following we will apply both one step ahead and four steps ahead forecasts. The four step ahead forecast sequence for each flexible regression method is constructed in a way analogous to the sequences constructed for the one step ahead forecast described above. The motivation for also considering four steps ahead forecasts for each flexible regression method is that linear models might locally approximate nonlinear patterns reasonably well. Hence, the one step-ahead forecast measure may not unveil the nonlinear components, while a four step ahead forecast could. A drawback is that the overall forecast ability of all the econometric models may fall dramatically as a consequence of extending the forecast horizon. Hence, we may end up with the difficult task of comparing forecasts which are all of a rather low quality.

The method we use here to produce the one and four steps ahead forecast is the so-called direct method. The primary reason for this particular choice is due to the conceptual and computational simplicity of the direct method. According to the direct method the h steps ahead forecasts are calculated simply as

$$\widehat{y}_{t+h} = \mu(x_t, \widehat{\varsigma}) \tag{76}$$

where ς is obtained from the general regression model $y_{t+h} = \mu(x_t, \varsigma) + \epsilon_t$ and hence will equal $a)$ the least square estimator of the linear regression coefficients in cases where $\mu(.)$ is a linear function in the regressors $b)$ the maximum likelihood estimator of φ when applied to Hamilton's Flexible Regression Model or $c)$ the nonlinear least square estimator of κ and ϱ in the Neural Network Regression Model or Projection Pursuit Regression Models, respectively.

3.2.5 Measures of the absolute forecast performance

Obviously the loss function applied in measuring forecast performance should be the economic loss associated with the produced forecast. However, such a loss cannot be computed here and consequently we apply several different measures, which enable us to evaluate how sensible our results may be to the choice of loss function. We compare the absolute predictive accuracy of the models using three different loss functions. The first measure is the forecast mean squared error (MSE), the second is the forecast mean absolute deviation (MAD), while the third is the forecast absolute percentage error (MAPE). Consider the h steps ahead forecast and let $j = 1, 2, ...T$ be the training or estimation sample period. If we let n denote the length of the out of sample period over which we wish to evaluate the forecast performance the three measures can be defined as

$$\begin{aligned} MSE &= \frac{1}{n}\sum_{j=1-h}^{n-h}(y_{T+h+j} - \widehat{y}_{T+h+j})^2 && (77) \\ MAD &= \frac{1}{n}\sum_{j=1-h}^{n-h}|y_{T+h+j} - \widehat{y}_{T+h+j}| && (78) \\ MAPE &= \frac{1}{n}\sum_{j=1-h}^{n-h}|\frac{\widehat{y}_{T+h+j}}{y_{T+h+j}} - 1| && (79) \end{aligned}$$

The forecast with the smallest value of these measures is producing the "best" forecast. Theil's U-statistic defined by

$$U = \left[\sum_{j=1-h}^{n-h}\frac{(y_{T+h+j} - \widehat{y}_{T+h+j})^2}{(y_{T+h+j} - y_{T+j})^2}\right]^{\frac{1}{2}} \tag{80}$$

is also a commonly applied measure. A value of $U = 0$ indicates a perfect forecast, while a value equal to 1 indicates that the forecast is equivalent to a no change forecast. Another measure of the absolute forecast performance is the squared correlation coefficient, R^2 , between the forecast and the actual value. However, the correlation coefficient does not measure whether the forecast is off target with a constant and therefore it can only be applied as a relevant measure if the intercept is zero and the slope equals one in a regression of the actual value on the forecast. The so-called Granger Newbold regression. Let $\{y_{T+h+j}\}_{j=1-h}^{n-h}$ and $\{\widehat{y}_{T+h+j}\}_{j=1-h}^{n-h}$ denote a sequence of actual and predicted values respectively. Consider the following linear regression

$$y_{T+h+j} = \alpha_0 + \alpha_1\widehat{y}_{T+h+j} + \epsilon_{T+h+j},\ j = 1 - h, .., n - h \tag{81}$$

The hypothesis of an unbiased out of sample forecast corresponds to $\alpha_0 = 1 - \alpha_1 = 0$. Thus a very simple regression based test is the student t test of the two separate null hypotheses $\mathrm{H}_0 : \alpha_0 = 0$ and $\mathrm{H}_0 : \alpha_1 = 1$. Furthermore, we apply the regression based F-test to test the composite hypothesis $\mathrm{H}_0 : \alpha_0 = 0$, $\alpha_1 = 1$. Finally, we compute the R^2 from the above regression and interpret it as a simple measure of absolute forecast performance.

3.2.6 Measures of directional forecast performance

In addition to the application of measures of the absolute forecast performance, measures of directional predictive ability are often used. Here we will consider the Henriksson-Merton test (HM), see Henriksson and Merton (1981), in order to compare the ability of the models under consideration to forecast the correct directions of the movements in the series y_t over time. The HM test for directional prediction ability is conducted by setting up the following 2 × 2 contingency table

		actual		
		up	down	
predicted	up	n_{11}	n_{12}	n_{10}
	down	n_{21}	n_{22}	n_{20}
		n_{01}	n_{02}	n

Assuming that the column and row sums are fixed the HM test can be interpreted as an exact test for independence between the predicted and the actual values of the time series y_t. The HM test statistics is calculated as

$$HM = \frac{n_{11} - \frac{n_{10}n_{01}}{n}}{\sqrt{\frac{n_{10}n_{01}n_{20}n_{02}}{n^2(n-1)}}} \sim N(0,1) \tag{82}$$

Asymptotically HM will have a standard normal distribution. However, as pointed our by Pesaran and Timmermann (1994) the HM test is difficult to interpret from an economic point of view since it is conditional on fixed margins where the margins are the predicted changes in downward direction (n_{10}), the realized changes in downward direction (n_{01}), and the realized changes in upward direction (n_{02}). In our case these quantities will be unknown and therefore they must to be estimated and consequently the test of independence looses two degrees of freedom. The HM test is still a relevant test statistics, but it is asymptotically equivalent to the well known χ^2 test for independence in a 2×2 contingency table given by

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - \frac{n_{i0}n_{0j}}{n})^2}{\frac{n_{i0}n_{0j}}{n}} \sim \chi^2(1) \tag{83}$$

Based on the 2 × 2 contingency table we will also present a measure of the confusion rate denoted CR which is the sum of the off diagonal elements divided by the total number of elements, and the degree of diagonal concentration denoted ϕ. The confusion rate CR and ϕ are computed as

$$CR = \frac{n_{12} + n_{21}}{n} \tag{84}$$

$$\phi = \sqrt{\chi^2/n} \tag{85}$$

The CR measure indicates how frequently the forecast is in the wrong direction, while the ϕ measure can be interpreted almost like an R^2, but only in the 2×2 case. The χ^2 statistics is proportional to n and in a 2×2 table ϕ will lie between zero, independence, and one, perfect directional prediction.

3.2.7 Application 1b. The change in the US Unemployment rate

Let us now proceed and evaluate the flexible regression models considered on the first differences in US unemployment rate and sequences of true one step ahead out of sample forecasts to evaluate each of the flexible regression models discussed in Section 2. In recent work on the search for nonlinear components in US macroeconomic time series, forecast comparisons between linear and flexible regression models have been done by the use of the exponential smoothing approach, see Stock and Watson (1998), the neural network model , see Swanson and White (1995, 1997a, 1997b) and Stock and Watson (1998), or the smooth transitions autoregressions., see Teräsvirta (1995) and Stock and Watson (1998). In case of forecast ability based on MSE in particular the evidence from these studies has been in favour of the linear model. However, as also pointed out by Swanson and White this could be due to the chosen model selection criterion and in fact they question the use of BIC for selecting the best model with respect to forecast performance.

The results of the model selection and forecasting exercise for the change in US unemployment rate are given in Tables 3, 4, and 5. In Table 3 all the models have been selected using the AIC criterion and the mean value of the AIC is shown in the first row. The second row gives the frequency of cases where the one step ahead forecast can be improved by adding the nonlinear component to the linear part. The computation of the frequency is done as follows. The best model according to the AIC criterion is specified for the period to and including 1979.Q4 and it is registered whether the best model contains the nonlinear part of the specification. In the Hamilton model this will be the part of equation(2) containing $\lambda m(g \odot x_t)$ while the linear part is $x_t^{'}\beta$. The model is used to forecast 1980.Q1.

Next, the best model according to the AIC criterion is specified for the period to and including 1980.Q1 and it is registered whether the best model contain the nonlinear part of the specification. The model is used to forecast 1980.Q2, etc. Measures of the absolute one period ahead forecast performance of each of the models are presented in the second row block of Table 3, while the row block contains the measures of the directional forecast performance. Tables 4, and 5 contain the corresponding information when the BIC and the CV criteria have been used to select the best model at each stage. Based on the results in Tables 3, 4, and 5, Hamilton's FNL model using the CV criterion performs the best when the absolute measures are applied, while Hamilton's FNL model using the BIC criterion is superior when the directional measures are applied.

The projection pursuit model PPR2, which is especially suitable in cases of only moderate nonlinearities, performs quite well especially when the model selection is based on the CV criterion and the directional measures are considered. In general, the AIC criterion is not producing the best forecast in any case. None of the selection criteria picked up a nonlinear component for the PPR2 model probably because it is very difficult to identify moderate nonlinearities. For that reason we imposed a nonlinear component in the PPR2 model by constraining v to obey $v \geq 1$ in equation (27). This also explains why the mean values of the various information criteria exceeds the ones associated with the linear models. While PPR2 always "finds" it necessary to apply the nonlinear component, PPR1 seldom does. The FNL model "finds" a nonlinear component necessary quite often if the BIC criterion is applied and a little more than half the times when the CV criterion is applied , but never when the AIC criterion is used The neural network model ANN applies a nonlinear component when the CV criterion is used in 91% of the cases but almost never or never when the other criteria are applied.

3.2.8 Application 2b. The growth rate of US industrial production

The results of the applications to the growth rate of the Industrial Production are presented in Tables 6, 7, and 8. The best performing models are here PPR2 applying the CV criterion when the absolute forecast performance measures are used and the ANN model applying the CV criterion when the directional forecast performance measures are used. That the FNL model does not perform as well as in the case of the change in the unemployment can be no surprise in the light of the result of the test for linearity performed earlier on, as Hamilton's test, HLM, did not reject linearity. Both the PPR2 model and the ANN model finds that a nonlinear component should be added in all cases. Hence, the general impression from the two applications performed here indicates that the model selected by the Cross Validation criterion in general has slightly better forecast abilities than the models selected by the BIC criterion. The models selected by the AIC criterion have the worst forecasting record. The best performing models with respect to the forecasting ability seem to be Hamilton's Flexible Regression model and the projection pursuit model for moderate non-linearities, PPR2, and the Neural Network Model ANN. The results indicate that the results obtained when testing for linearity are consistent with the results applied when estimating the full model, insofar as both the result of the HLM test and the result of the $NNLM$ test are consistent with the result obtained when estimating the model by FNL and ANN, respectively.

3.3 Comparing real time forecast abilities

The best forecasting model specification from the four classes of nonlinear models is compared to the linear model by use of the Diebold-Mariano test and a series of forecasting encompassing tests including those suggested by Harvey et al. Again the procedures are applied to the two series: The change in US unemployment rate and the growth rate of US industrial production.

3.3.1 The Diebold-Mariano test

Comparing the relative forecast accuracy of two different time series models is complicated for a number of reasons. Forecasts will typically be serially and contemporaneously correlated and will often be characterized by having a non-Gaussian heavy tailed distribution. Fortunately, the test of the null hypothesis of no difference in the accuracy of two competing forecasts suggested by Diebold and Mariano (1995) (DM) is robust against all of these features. The DM test statistics is designed as follows: Let $\{\widehat{y}_{it}\}_t^n$ and $\{\widehat{y}_{jt}\}_t^n$ denote two sequences of forecasts of the time series $\{y_t\}_t^n$ and let the associated forecast errors be $\{e_{it}\}_t^n$ and $\{e_{jt}\}_t^n$. Assume that the loss function[12], $g(.)$, can be written as a function of the forecast errors only i.e. $g(e_{it})$ and define the loss differential between the two competing forecasts as $d_t \equiv [g(e_{it}) - g(e_{jt})]$. If the sequence $\{d_t\}_t^n$ is covariance stationary and has a short memory the asymptotic distribution of the sample mean loss differential $\bar{d} = (1/n)\sum_{t=1}^n d_t$ can be shown to be, see Diebold and Mariano (1995),

$$\sqrt{n}(\overline{d} - \mu) \xrightarrow{d} N(0, V(\overline{d})) \tag{86}$$

If all autocorrelations of order h or higher of the sequence $\{d_t\}_t^t$ are all zero when considering h steps ahead forecasts, the variance $V(\overline{d})$ can be shown to be equal to

$$V(\overline{d}) = n^{-1}(r_0 + 2\sum_{i=1}^{h-1} r_i) \tag{87}$$

where r_i is the i'th order autocorrelation of d_t. An estimate of r_i can be obtained as

$$\widehat{r}_i = n^{-1}\sum_{t=i+1}^{n}(d_t - \overline{d})(d_{t-i} - \overline{d}) \tag{88}$$

and by substituting the estimate for $\widehat{r}_i$ in the expression for $V(\bar{d})$ we obtain the variance estimate $\widehat{V}(\overline{d})$. The Diebold-Mariano test statistic for testing the null hypothesis of equal forecast accuracy is then given by

$$DM = \frac{\overline{d}}{\sqrt{\widehat{V}(\overline{d})}} \sim N(0,1) \tag{89}$$

[12] Results applying the MSE loss function will be reported below.

Under the null hypothesis the Diebold-Mariano test statistics will have a standard normal distribution asymptotically. Harvey, Leybourne and Newbold (1997) argue that the DM test can be quite over-sized in small and moderate samples and that this problem becomes acute for longer forecast horizons. In order to cure this problem they suggest modifying the DM statistics as follows

$$MDM = \sqrt{\frac{n+1-2h+n^{-1}h(h-1)}{n}} * DM \tag{90}$$

Furthermore Harvey, Leybourne and Newbold (1997) suggest comparing the statistics with critical values from the Student's t distribution with $n-1$ degrees of freedom rather than from the standard normal distribution.

3.3.2 The forecast encompassing tests.

Evaluation of the relative forecast ability of two competing econometric models can also be done by using the forecast encompassing principle, see Chong and Hendry (1986), Clements and Hendry (1993) , and Harvey, Leybourne and Newbold (1998). The forecast encompassing principle is based on the literature on combining forecast, see Granger and Newbold (1977). Consider the combined forecast y_{ct} from two competing forecasts given as

$$\begin{aligned} y_{ct} &= (1-\alpha)\widehat{y}_{it} + \alpha\widehat{y}_{jt} && (91) \\ 0 &\leq \alpha \leq 1 && (92) \end{aligned}$$

Then the model generating the forecast $\widehat{y}_{it}$ is said to forecast encompass the other model i.e. $\widehat{y}_{jt}$ if the entire optimal weight is associated with $\widehat{y}_{it}$t, that is if $\alpha = 0$. If we define $\epsilon_t = y_t - y_{ct}$ the equation above can be rewritten as

$$e_{it} = \alpha(e_{it} - e_{jt}) + \epsilon_t \tag{93}$$

This implies that given the sequence of forecast errors $\{e_{it}\}_t^n$ and $\{e_{jt}\}_t^n$ we can determine whether $\widehat{y}_{it}$ forecast encompasses $\widehat{y}_{jt}$ by applying standard regression based tests of the null hypothesis $\alpha = 0$. Furthermore, the test procedure is expected to perform well when (e_{it}, e_{jt}) is bivariate normally distributed. However, this is probably very rarely the case. In addition, Harvey, Leybourne and Newbold (1998) show analytically and by Monte Carlo simulations that if the distribution of (e_{it}, e_{jt}) deviates from the bivariate normal the test for forecast encompassing can be seriously oversized.

A possibility for handling deviations from the assumption of bivariate normality of the variables in equation (93), is to apply the rank correlation between e_{it} and $e_{it} - e_{jt}$ as the basis for the test of $\alpha = 0$. The rank correlation is obtained by ordering the "observations" for e_{it} and $e_{it} - e_{jt}$ according to size and measure the relationship between their ranks instead of their actual numerical values. Spearman's rank correlation coefficient is given by

$$R_s = 1 - \frac{6\sum_{\tau} DR_{\tau}^2}{n(n^2-1)} \tag{94}$$

with the approximate distribution

$$(n-1)^{\frac{1}{2}}R_s \sim N(0,1) \tag{95}$$

where DR_τ equals the difference between ranks of corresponding pairs of e_{it} and $e_{it}-e_{jt}$ and n equals the number of "observations". The approximation is known to be reasonably good provided $n > 30$. A small value of R_s corresponds to a small value of α. Hence a small value of R_s indicates that the forecast $\widehat{y}_{jt}$ does not contribute to the combined forecast and consequently $\widehat{y}_{it}$ encompasses $\widehat{y}_{jt}$. Similarly, whether $\widehat{y}_{jt}$ encompasses $\widehat{y}_{it}$ may be tested by help of Spearman's rank correlation coefficient between e_{jt} and $e_{jt}-e_{it}$.

Even if application of Spearman's' rank correlation statistics solves the problem of nonnormality of the included variables it cannot guard against dependence of the errors in the auxiliary regression in (93). Harvey, Leybourne and Newbold (1998) therefore suggest robustifying the forecast encompassing test against such dependence. The robustified tests are based on the assumption that the h steps ahead forecasts at most have forecast errors that are $h-1$ dependent. If this is correct a robust forecast encompassing test statistics can be calculated as

$$R_\upsilon = n^{1/2}\widehat{Q}_\upsilon^{-1/2}\overline{D} \sim N(0,1), \upsilon = 1,2 \tag{96}$$

where $D_t = (e_{it}-e_{jt})e_{it}$, and $\overline{D} = n^{-1}\sum_{t=1}^{n} D_t$. If we let $\widehat{\epsilon_t}$ be the residual from a least square regression of e_{it} on $(e_{it}-e_{jt})$ the estimator of Q_1 can be obtained as

$$\widehat{Q}_1 = n^{1/2}\sum_{\tau=-(h.-1)}^{h-1}\sum_{t=|\tau|+1}^{n}(e_{it}-e_{jt})\widehat{\epsilon}_t(e_{i,t-|\tau|}-e_{j,t-|\tau|})\widehat{\epsilon}_{t-|\tau|} \tag{97}$$

The estimator $\widehat{Q}_1$ is a consistent estimator but convergence may be slow. Instead an alternative estimator denoted $\widehat{Q}_2$ could be applied. The estimator of Q_2 can be computed as

$$\widehat{Q}_2 = n^{1/2}\sum_{\tau=-(h.-1)}^{h-1}\sum_{t=|\tau|+1}^{n}D_tD_{t-|\tau|} \tag{98}$$

Unfortunately, $\widehat{Q}_2$ is a consistent estimator only under the null. Asymptotically both R_1 and R_2 have a standard normal distribution under the null, although Monte Carlo evidence, see Harvey, Leybourne and Newbold (1998), suggests that the tests have better size and power properties when using the critical values from the Students t distribution with $n-1$ degrees of freedom. However, R_2 may have a lack of power due to the use of $\widehat{Q}_2$. The tests R_2 may also be improved by replacing $\widehat{Q}_2$ by

$$\widehat{Q}_{dm} = n^{1/2}\sum_{\tau=-(h.-1)}^{h-1}\sum_{t=|\tau|+1}^{n}[D_t-\overline{D}][D_{t-|\tau|}-\overline{D}] \tag{99}$$

i.e. by applying the mean corrected D_t The improved test statistics is denoted R_{dm}. Following Harvey, Leybourne and Newbold (1998) R_{dm} may be improved even further by a correction similar to the one applied to DM above. Hence, the test statistics becomes

$$R_{mdm} = \sqrt{\frac{n+1-2h+n^{-1}h(h-1)}{n}} R_{dm}. \tag{100}$$

In order to save space only forecast comparisons between the best linear model and the best non-linear model are reported. In determining the best models the measures applied are the absolute forecast performance measures $MSE, MAD, MAPE$, and R^2 and the directional forecast performance measure ϕ depicted in Tables 9 and 10. Both the 1 step ahead forecast ability and the 4 steps ahead forecast ability are compared. The motivation for also considering four steps ahead forecasts for each flexible regression method is that linear models might approximate nonlinear patterns reasonably well when the forecast horizon is small. When performing the 4 steps ahead forecast comparisons the R and the R$_s$ statistics are not used due to their sensitivity to autocorrelation in the errors.

3.3.3 Application 1c. The change in the US unemployment rate

The Diebold-Mariano test results, DM and the Modified DM test results shown in Table 9 indicate that the non-linear models are no better in predicting the change in the US unemployment than the linear model. A different result is obtained when the encompassing test is applied, see Table 10. When the absolute performance measures are used in selecting the best models, the best non-linear model, FNL^{CV} encompasses the best linear model and **not** vice versa in the 1 step ahead forecasts. This result is obtained irrespective of the choice of test statistic R, R_s, R_1, R_{dm}, and R_{mdm}. However, the best nonlinear model PPR2^{CV} found when the directional performance measure ϕ is applied, encompasses the best linear model and vice versa in the 1 step ahead forecast of the change in the unemployment. The results for the 4 steps ahead forecasts are somewhat more mixed. When the absolute performance measures MSE, MAD and the directional measure ϕ are applied, PPR2 using the CV criterion, i.e. PPR2^{CV} is the preferred nonlinear model, and it encompasses the best linear model and **not** vice versa at least at a 10 % level of significance. However, when the measure applied is the $MAPE$ and R^2, the best nonlinear model FNL^{CV} and the best linear model LR^{CV} encompass each other.

3.3.4 Application 2c.The growth rate of US industrial production

Again the Diebold-Mariano test results, DM and the Modified DM test results shown in Table 11 indicate that the non-linear models are no better in predicting the growth rate of US industrial production than the preferred linear model. However, the forecast encompassing tests, see Table 12 indicate that the best nonlinear model encompasses the linear model but **not** vice versa. This result is obtained irrespective of whether 1 step ahead or 4 steps ahead forecast are applied or whatever test is applied except Spearman's rank correlation test R_s. When the directional measure ϕ is applied the result is less clear cut in the 1 step ahead forecast comparisons. In the 1 step ahead forecast PPR2^{CV,BIC} are the best models when the absolute forecast measures are applied, while ANN^{CV} is the preferable model when the directional measures are applied. But irrespective of the result of Hamilton's test for linearity, which indicates that the model is linear, FNL^{BIC,CV} are the best models when 4 steps ahead forecast are performed for the growth rate of US industrial production.

Hence for both series investigated here, i.e. the change in the US unemployment rate and the growth rate of US industrial production, there is some indication that the non-linear model is preferable, and that FNL and PPR2 are the better ones. In addition, the results also suggest that the small sample power properties of the forecast encompassing tests are better than the small sample properties of the Diebold Mariano test, at least when the modelling procedure starts from the linear specification and adds the nonlinear parts. In addition, the size properties of the encompassing tests seem quite acceptable, as shown by Harvey et. al. (1998).

4 Conclusions

From the limited empirical evidence obtained here it is tentatively suggested to find a baseline nonlinear flexible form for a univariate time series by following the procedure: **1**. Recursively, based on h extra periods at a time specify and estimate a linear form by use of model selection criteria like Cross Validation and/or BIC. **2**. After a preliminary test for linearity, recursively, specify and estimate flexible regression models like the FNL suggested by Hamilton (1999) and the Projection Pursuit model suggested by Aldrin, Boelviken and Schweder (1993) for cases of moderate nonlinearities. Use the Cross Validation and the BIC criteria. **3**. Based on the remaining part of the data set select the best nonlinear flexible form by use of forecast criteria measuring the absolute forecast performance and the directional forecast performance in h-steps ahead predictions, and compare the best flexible form to the linear specification by use of the Diebold Mariano tests, see Diebold and Mariano (1995) and the forecast encompassing tests suggested by Harvey, Leybourne, and Newbold (1998). The results indicate that the FNL method and the Projection Pursuit Model are the preferable models to apply and that the CV and BIC are the best selection criteria, while the forecast encompassing tests properly modified as suggested by Harvey et. al.(1998) possess better power properties than the Diebold-Mariano test.

References

Akaike, H. (1969), 'Fitting autoregressions for prediction', *Annals of the Institute of Statistical Mathematics* **21**, 243–247.

Aldrin, M., Boelviken, E. & Schweder, T. (1993), 'Projection pursuit regression for moderate non-linearities', *Computational Statistics and Data Analysis* **16**, 379–403.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, New York.

Chong, Y. & Hendry, D. (1986), 'Econometric evaluation of linear macroeconomic models', *Review of Economic Studies* **53**, 671–690.

Clements, M. & Hendry, D. (1993), 'On the limitations of comparing mean squared forecast errors', *Journal of Forecasting* **12**, 617–637.

Dahl, C. M. (1999), 'An investigation of tests for linearity and the accuracy of flexible nonlinear inference'. Unpublished manuscript, Department of Economics, University of Aarhus.

Diebold, F. & Mariano, R. (1995), 'Comparing predictive accuracy', *Journal of Business and Economic Statistics* **13**, 253–263.

Friedman, J. & Stueltze, W. (1981), 'Projection pursuit regression', *Journal of American Statistical Association* **76**, 817–823.

Granger, C. W. J. & Newbold, P. (1977), *Forecasting Economic Time Series*, Academic Press, New York.

Granger, C. W. J. & Teraesvirta, T. (1993), *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford, New York.

Haerdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, New York, NY.

Hamilton, J. D. (1999), 'A parametric approach to flexible nonlinear inference'. Unpublished manuscript, Department of Economics, University of California, San Diego.

Harvey, D., Leybourne, S. & Newbold, P. (1997), 'Testing the equality of prediction mean squared errors', *International Journal of Forecasting* **13**, 281–291.

Harvey, D., Leybourne, S. & Newbold, P. (1998), 'Tests for forecast encompassing', *Journal of Business and Economic Statistics* **16**, 254–259.

Henriksson, R. & Merton, R. (1981), 'On market timing and investment performance. ii. statistical procedures for evaluating forecasting skills', *Journal of Business* **54**, 513–533.

Huber, P. (1985), ‘Projection pursuit’, *Annals of Statistics* **13**, 435–475.

Lee, T.-H., White, H. & Granger, C. W. J. (1993), ‘Testing for neglected nonlinearity in time series models’, *Journal of Econometrics* **56**, 269–290.

Pesaran, M. & Timmermann, A. (1994), ‘A generalization of the non-parametric henriksson-merton test of market timing’, *Economic Letters* **44**, 1–7.

Ramsey, J. B. (1969), ‘Tests for specification errors in classical linear least square regression analysis’, *Journal of the Royal Statistical Society* **Series B, 31**, 350–371.

Schwarts, G. (1978), ‘Estimating the dimension of a model’, *Annals of Statistics* **6**, 461–464.

Stock, J. H. & Watson, M. W. (1998), ‘A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series’, *NBER, Working Paper* (6607).

Stone, C. (1977), ‘Consistent nonparametric regression (with discussion)’, *Annals of Statistics* **5**, 595–645.

Stone, M. (1974), ‘Cross-validatory choice and assessment of statistical predictions (with discussion)’, *Journal of Royal Statistical Society, Series B* **36**, 111–147.

Swanson, N. R. & White, H. (1995), ‘A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks’, *Journal of Business and Economic statistics* **13**(3), 265–275.

Swanson, N. R. & White, H. (1997*a*), ‘Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models’, *International Journal of Forecasting* (13), 439–461.

Swanson, N. R. & White, H. (1997*b*), ‘A model selection approach to realtime macroeconomic forecasting using linear models and artificial neural networks’, *The Review of Economics and Statistics* pp. 541–550.

Teraesvirta, T. (1995), ‘Modelling nonlinearity in u.s. gross national product 1889-1987’, *Empirical Economics* **20**, 577–597.

Tsay, R. S. (1986), ‘Nonlinearity test for time series’, *Biometrica* **73**, 461–466.

Wahba, G. (1977), A survey of some smoothing problems and the method of generalized cross-validation for solving them, *in* P. Krishnaiah, ed., ‘Application of Statistics’, Amsterdam, North Holland.

Wahba, G. & Wold, S. (1975), ‘A completely automatic french curve: fitting spline functione by cross validation’, *Communications in Statistics, Series A* **4**, 1–17.

Wecker, W. E. & Ansley, C. F. (1983), ‘The signal extraction approach to nonlinear regression and spline smoothing’, *Journal of the American Statistical Association* **78**, 81–89.

White, H. (1989), An additional hidden unit test for neglected nonlinearity in multilayer feedforward networks, *in* ‘Proceedings of the international joint conference on neural networks’, IEEE Press, New York, NY, Washington, DC, pp. 451–455.

White, H. (1992), *Estimation, Inference and Specification Analysis*, Cambridge University Press, New York, NY.

5 Tables

Table 1: Tests for linearity. Change in US Unemployment rate, 1949.Q3 - 1998.Q2

TEST	AIC (lag=4)	BIC (lag=2)	CV (lag=4)
HLM	7.18 (0.01)	5.65 (0.02)	7.18 (0.01)
$NNLM$	10.16 (0.00)	15.61 (0.00)	10.16 (0.00)
$TSAYF$	3.32 (0.00)	4.51 (0.00)	3.32 (0.00)
WIM	32.29 (0.03)	17.11 (0.03)	32.29 (0.03)
$RESETF$	3.54 (0.06)	3.05 (0.08)	3.54 (0.06)

$*$ p-values in parentheses. HLM-Hamilton's LM test
$NNLM-$The Neural Network test. $TSAYF-$The Tsay test
$WIM-$ White's Inf. Matrix test. $RESETF-$ The RESET test

Table 2: Tests for linearity. Growth rate in US Industrial Production, 1947.Q2-1998.Q2

Test	AIC (lag=2)	BIC (lag=2)	CV (lag=2)
HLM	0.75 (0.39)	0.75 (0.39)	0.75 (0.39)
$NNLM$	11.39 (0.00)	11.39 (0.00)	11.39 (0.00)
$TSAYF$	2.77 (0.04)	2.77 (0.04)	2.77 (0.04)
WIM	17.00 (0.03)	17.00 (0.03)	17.00 (0.03)
$RESETF$	2.03 (0.15)	2.03 (0.15)	2.03 (0.15)

$*$see the note to Table 1

Table 3: One period ahead forecast performance of the change in US Unemployment rate, 1980.Q1 - 1998.Q2. All models selected by AIC

	LR	FNL	ANN	PPR1	PPR2
Mean AIC	-2.3014	-2.3014	-2.3095	-2.3891	-2.2588
Frequency *		0.00	0.09	0.22	1.00
Absolute forecast . performance					
MSE	0.0679	0.0679	0.0758	0.1196	0.0682
MAD	0.1950	0.1950	0.2020	0.2614	0.1949
$MAPE$	2.0852	2.0852	2.1093	3.4814	2.0409
Theils U	0.9621	0.9621	1.0161	1.2764	0.9640
t-stat.(intc=0) [p-val.]	0.4650	0.4650	0.4274	0.1979	0.3718
t-stat.(slope=1) [p-val.]	0.5684	0.5684	0.1937	0.0002	0.4005
F-statistic [p-val.]	0.6512	0.6512	0.3173	0.0002	0.4669
R^2	0.3600	0.3600	0.3009	0.1603	0.3636
Directional forecast performance					
HM [p-val.]	0.2552	0.2552	0.2552	0.5134	0.1657
χ^2 [p-val.]	0.2520	0.2520	0.2520	0.5106	0.1629
CR	0.4324	0.4324	0.4324	0.4595	0.4189
ϕ	0.1332	0.1332	0.1332	0.0765	0.1622

* The frequency of a nonlinear component being added

**LR-Linear Regression Model, FNL-Hamilton's Flexible R.M ANN- The Neural Network R.M., PPR1-The Projection Pursuit R.M., PPR2- The Projection Pursuit R.M. for Moderate N.L.

Table 4: One period ahead forecast performance of the change in US Unemployment rate, 1980.Q1 - 1998.Q2. All models selected by BIC

	LR	FNL	ANN	PPR1	PPR2
Mean BIC	-2.2297	-5.4636	-2.2297	-2.2348	-2.1636
Frequency		0.88	0.00	0.22	1.00
Absolute forecast performance					
MSE	0.0668	0.0897	0.0668	0.0667	0.0675
MAD	0.1950	0.2144	0.1950	0.1968	0.1973
$MAPE$	2.7775	3.0734	2.7775	2.8815	2.8723
Theil's U	0.9543	1.1054	0.9543	0.9539	0.9594
t-stat.(intc=0) [p-val.]	0.6841	0.1994	0.6841	0.7452	0.5320
t-stat.(slope=1) [p-val.]	0.4693	0.0001	0.4693	0.3461	0.2066
F-statistic [p-val.]	0.7183	0.0007	0.7183	0.6205	0.3827
R^2	0.3685	0.3378	0.3685	0.3716	0.3733
Directional forecast performance					
HM [p-val.]	0.2573	0.0181	0.2573	0.5085	0.5132
χ^2 [p-val.]	0.2540	0.0174	0.2540	0.5057	0.5103
CR	0.4324	0.3649	0.4324	0.4594	0.4595
ϕ	0.1326	0.2765	0.1326	0.0774	0.0765

* See the notes to Table 3

Table 5: One period ahead forecast performance of the change in US Unemployment rate, 1980.Q1 - 1998.Q2. All models selected by CV

	LR	FNL	ANN	PPR1	PPR2
Mean CV	0.1079	0.1056	0.0969	0.1075	0.1110
Frequency		0.58	0.91	0.08	1.00
Absolute forecast performance					
MSE	0.0669	0.0631	0.0665	0.0665	0.0670
MAD	0.1932	0.1895	0.1951	0.1939	0.1935
$MAPE$	2.0614	1.9739	2.0537	2.0848	2.1403
Theil's U	0.9548	0.9275	0.9522	0.9520	0.9556
t-stat.(intc=0) [p-val.]	0.5118	0.5298	0.4080	0.5642	0.5775
t-stat.(slope=1) [p-val.]	0.6309	0.6715	0.5638	0.6315	0.3304
F-statistic	0.7208	0.7532	0.5986	0.7592	0.5437
R^2	0.3679	0.4028	0.3745	0.3706	0.3719
Directional forecast performance					
HM [p-val.]	0.1692	0.0665	0.1692	0.2572	0.0206
χ^2 [p-val.]	0.1663	0.0647	0.1663	0.2540	0.0198
CR	0.4189	0.3919	0.4189	0.4324	0.3649
ϕ	0.1609	0.2148	0.1609	0.1326	0.2709

* See the notes to Table 3

Table 6: One period ahead forecast performance of the growth rate of US Industrial Production, 1980.Q1 - 1998.Q2. All models selected by AIC

	LR	FNL	ANN	PPR1	PPR2
Mean AIC	1.4509	-2.2340	1.4336	1.4101	1.4874
Frequency		0.77	0.95	0.68	1.00
Absolute forecast performance					
MSE	1.3619	1.5388	1.4778	1.4779	1.3808
MAD	0.7960	0.9040	0.8628	0.8733	0.8117
$MAPE$	1.1240	1.1768	1.3322	1.3413	1.2314
Theil's U	0.8846	0.9403	0.9215	0.8803	0.8908
t-stat.(intc=0)[p-val.]	0.4366	0.5747	0.5505	0.8290	0.9885
t-stat.(slope=1)[p-val.]	0.8964	0.0553	0.1122	0.1708	0.2851
F-stat [p-val.]	0.3897	0.1091	0.2458	0.2329	0.3322
R^2	0.2338	0.1671	0.1799	0.1812	0.2268
Directional forecast performance					
HM [p-val.]	0.0072	0.0188	0.2605	0.2043	0.1145
χ^2 [p-val.]	0.0068	0.0180	0.2573	0.2013	0.1121
CR	0.3378	0.3649	0.4459	0.4324	0.4054
ϕ	0.3144	0.2749	0.1317	0.1486	0.1847

* See the notes to Table 3

Table 7: One period ahead forecast performance of the growth rate of US Industrial Production, 1980.Q1 - 1998.Q2. All models selected by BIC

	LR	FNL	ANN	PPR1	PPR2
Mean BIC	1.5076	0.7740	1.5076	1.5068	1.7640
Frequency		0.42	0.00	0.51	1.00
Absolute forecast performance					
MSE	1.3619	1.7928	1.3619	1.4611	1.3094
MAD	0.7960	0.9345	0.7960	0.8377	0.7893
$MAPE$	1.1240	1.8139	1.1240	1.0860	1.1123
Theil's U	0.8846	1.0150	0.8846	0.9163	0.8675
t-stat.(intc=0)[p-val.]	0.4366	0.1630	0.4366	0.9476	0.2328
t-stat.(slope=1)[p-val.]	0.8964	0.0043	0.8964	0.3876	0.9287
F-stat.[p-val.]	0.3897	0.0209	0.3897	0.4877	0.2141
R^2	0.2338	0.0815	0.2338	0.1725	0.2762
Directional forecast performance					
HM [p-val.]	0.0072	0.1405	0.0072	0.2871	0.0070
χ^2 [p-val.]	0.0068	0.1378	0.0068	0.2839	0.0074
CR	0.3378	0.4189	0.3378	0.4459	0.3378
ϕ	0.3144	0.1725	0.3144	0.1246	0.3112

* See the notes to Table 3

Table 8: One period ahead forecast performance of the growth rate of US Industrial Production, 1980.Q1 - 1998.Q2. All models selected by CV

	LR	FNL	ANN	PPR1	PPR2
Mean CV	4.5321	4.5321	4.2569	4.5273	4.5809
Frequency		0.00	1.00	0.05	1.00
Absolute forecast . performance					
MSE	1.3619	1.3619	1.3347	1.3614	1.2985
MAD	0.7960	0.7960	0.7965	0.7958	0.7791
$MAPE$	1.1240	1.1240	1.1371	1.1236	1.0785
Theil's U	0.8846	0.8846	0.8758	0.8846	0.8638
t-stat.(intc=0)[p-val.]	0.4366	0.4366	0.2736	0.4305	0.5479
t-stat.(slope=1)[p-val.]	0.8964	0.8964	0.9880	0.8845	0.9136
F-stat [p-val.]	0.3897	0.3897	0.2348	0.3718	0.6041
R^2	0.2338	0.2338	0.2603	0.2351	0.2601
Directional forecast performance					
HM [p-val.]	0.0072	0.0072	0.0039	0.0072	0.0215
χ^2 [p-val.]	0.0068	0.0068	0.0037	0.0068	0.0207
CR	0.3378	0.3378	0.3243	0.3378	0.3649
ϕ	0.3144	0.3144	0.3379	0.3144	0.2696

* See the notes to Table 3

Table 9: Diebold-Mariano test for relative predictive ability. Change in US Unemployment rate, 1980.Q1 - 1998.Q2. The Squared Error loss function is used in the DM and MDM statistics.

Measure	H_0	DM	MDM
	1 step ahead forecast		
MSE, R^2	$\mathrm{LR}^{BIC} \simeq \mathrm{FNL}^{CV}$	-0.938	-0.932
		(0.35)	(0.35)
$MAD, MAPE, \phi$	$\mathrm{LR}^{CV} \simeq \mathrm{FNL}^{CV}$	-0.691	-0.686
		(0.49)	(0.50)
	4 steps ahead forecasts		
MSE, ϕ	$\mathrm{LR}^{AIC} \simeq \mathrm{PPR2}^{BIC}$	-0.502	-0.837
		(0.62)	(0.41)
MAD	$\mathrm{LR}^{CV} \simeq \mathrm{PPR2}^{BIC}$	-0.006	-0.006
		(1.00)	(1.00)
$MAPE, R^2$	$\mathrm{LR}^{CV} \simeq \mathrm{FNL}^{CV}$	-0.929	-0.923
		(0.35)	(0.36)

* p-values in parentheses

$\simeq$ - Equal forecast accuracy

DM - The Diebold-Mariano test

MDM - The modified Diebold-Mariano test

Table 10: Encompassing test for relative predictive ability. Change in US Unemployment rate, 1980.Q1 - 1998.Q2. The best linear model versus the best non-linear model

Meas.	H_0	R	R_s	R_1	R_{dm}	R_{mdm}
				1 step ahead forecasts		
MSE,	LR^{BIC} $\sqsubset$	2.074	2.120	2.073	1.802	1.790
R^2	FNL^{CV}	(0.04)	(0.03)	(0.04)	(0.07)	(0.07)
	FNL^{CV} $\sqsubset$	0.117	-0.001	0.117	0.117	0.116
	LR^{BIC}	(0.91)	(0.91)	(0.91)	(0.91)	(0.91)
MAD,	LR^{CV} $\sqsubset$	2.129	-1.909	2.168	1.880	1.867
$MAPE$	FNL^{CV}	(0.03)	(0.06)	(0.03)	(0.06)	(0.06)
	FNL^{CV} $\sqsubset$	-0.406	-1.795	-0.414	-0.407	-0.405
	LR^{CV}	(0.68)	(0.07)	(0.68)	(0.68)	(0.69)
ϕ	LR^{CV} $\sqsubset$	(0.68)	0.193	0.661	0.633	0.629
	PPR2^{CV}	(0.51)	(0.85)	(0.51)	(0.53)	(0.53)
	PPR2^{CV} $\sqsubset$	0.754	1.334	0.749	0.743	0.738
	LR^{CV}	(0.45)	(0.18)	(0.45)	(0.46)	(0.46)
				4 steps ahead forecasts		
MSE,	LR^{AIC} $\sqsubset$	-	-	1.982	1.473	1.463
ϕ	PPR2^{BIC}	-	-	(0.05)	(0.14)	(0.14)
	PPR2^{BIC} $\sqsubset$	-	-	0.908	0.936	0.930
	LR^{AIC}	-	-	(0.36)	(0.35)	(0.35)
MAD	LR^{CV} $\sqsubset$	-	-	2.221	1.708	1.696
	PPR2^{BIC}	-	-	(0.03)	(0.09)	(0.09)
	PPR2^{BIC} $\sqsubset$	-	-	0.844	0.851	0.845
	LR^{CV}	-	-	(0.40)	(0.39)	(0.40)
$MAPE$,	LR^{CV} $\sqsubset$	-	-	1.330	1.301	10.292
R^2	FNL^{CV}	-	-	(0.18)	(0.19)	(0.20)
	FNL^{CV} $\sqsubset$	-	-	1.490	1.441	1.432
	LR^{CV}	-	-	(0.14)	(0.15)	(0.15)

* p-values in parentheses. A $\sqsubset$ B: A encompass B.

R - OLS based t-test.

R_s - Spearman's Rank correllation test, R_1 - Robust Encompassing test.

R_{dm} - Robust Encom. test with mean correct.

R_{mdm} - Robust Encom. test with mean and D.F. correct.

Table 11: Diebold-Mariano test for relative predictive ability. Growth rate of US Industrial Production. 1980.Q1 - 1998.Q2. The Squared Error loss function is used in the DM and MDM statistics.

Measure	H_0	DM	MDM
	1 step ahead forecasts		
$MSE, MAD, MAPE$	$\text{LR}^{CV} \simeq \text{PPR2}^{CV}$	-0.938	-0.932
		(0.35)	(0.35)
ϕ	$\text{LR}^{CV} \simeq \text{FNL}^{CV}$	0.000	0.000
		(1.00)	(1.00)
R^2	$\text{LR}^{BIC} \simeq \text{PPR2}^{BIC}$	-1.016	-1.010
		(0.31)	(0.32)
	4 steps ahead forecasts		
$MSE, MAPE, R^2$	$\text{LR}^{CV} \simeq \text{FNL}^{BIC}$	-0.732	-0.697
		(0.46)	(0.49)
MAD, ϕ	$\text{LR}^{CV} \simeq \text{FNL}^{CV}$	-1.733	-1.651
		(0.08)	(0.10)

* See the note to Table 9

Table 12: Encompassing test for relative predictive ability. Growth rate of US Industrial Production, 1980.Q1 - 1998.Q2. The best linear model versus the best non-linear model

Meas.	H_0	R	R_s	R_1	R_{dm}	R_{mdm}
		1 step ahead forecasts				
$MSE, MAD,$	LR^{CV} $\sqsubset$	2.204	1.061	1.978	1.793	1.781
$MAPE$	PPR2^{CV}	(0.03)	(0.29)	(0.05)	(0.07)	(0.07)
	PPR2^{CV} $\sqsubset$	-1.109	0.847	-0.995	-0.957	-0.951
	LR^{CV}	(.27)	(.0.40)	(0.32)	(0.34)	(0.34)
ϕ	LR^{CV} $\sqsubset$	1.259	1.580	2.279	1.337	1.320
	ANN^{CV}	(0.21)	(0.11)	(0.02)	(0.18)	(0.18)
	ANN^{CV} $\sqsubset$	0.309	1.407	0.576	0.527	0.530
	LR^{CV}	(0.76)	(0.16)	(0.58)	(0.53)	(0.53)
R^2	LR^{CV} $\sqsubset$	1.710	-0.688	1.905	1.945	1.932
	PPR2^{BIC}	(0.09)	(0.49)	(0.06)	(0.05)	(0.05)
	PPR2^{BIC} $\sqsubset$	-0.045	-1.241	-0.050	-0.050	-0.050
	LR^{CV}	(0.96)	(0.21)	(0.96)	(0.96)	(0.96)
		4 steps ahead forecasts				
$MSE,$	LR^{CV}$\sqsubset$	-	-	2.634	2.245	2.139
$MAPE, R^2$	FNL^{BIC}	-	-	(0.01)	(0.02)	(0.03)
	FNL^{BIC}$\sqsubset$	-	-	1.005	1.005	0.958
	LR^{CV}	-	-	(0.31)	(0.31)	(0.34)
MAD, ϕ	LR^{CV}$\sqsubset$	-	-	4.046	2.524	2.405
	FNL^{CV}	-	-	(0.00)	(0.01)	(0.02)
	FNL^{CV}$\sqsubset$	-	-	-0.054	-0.054	-0.051
	LR^{CV}	-	-	(0.96)	(0.96)	(0.96)

* See the note to Table 10

Working Paper

1998-11 Robert R. Dogonowski: Should Federal or Regional Insurance Protect the EMU?

1998-12 Robert R. Dogonowski: Income Taxation, Imperfect Competition and the Balanced Budget Multiplier.

1998-13 Ebbe Yndgaard: EU Enlargement - Latvia is Ready.

1998-14 Torben M. Andersen: Staggered Wage-Setting and Output Persistence.

1998-15 Henning Bunzel, Bent Jesper Christensen, Niels Haldrup, Svend Hylleberg, Viggo Høst, Peter Jensen og Allan Würtz: Udviklingslinier i Økonometrien.

1998-16 Peter Skott: Wage Formation and the (Non-)Existence of the NAIRU.

1998-17 Svend Hylleberg and Per Baltzer Overgaard: Competition Policy with a Coasian Prior?

1998-18 Lykke Eg Andersen and Osvaldo Nina: Micro-Credit and Group Lending: The Collateral Effect.

1998-19 Torben M. Andersen and Svend Hylleberg: Sources of Persistence in Employment Adjustment - Denmark 1974-1993.

1999-1 Bo Sandemann Rasmussen: Balancing the Budget: Long-Run Adverse Effects of Progressive Taxation.

1999-2 Palle Schelde Andersen, Mark Klau, and Ebbe Yndgaard: Higher Profits and Lower Capital Prices: Is Factor Allocation Optimal?

1999-3 Niels Haldrup and Michael Jansson: Spurious Regression, Cointegration, and Near Cointegration: A Unifying Approach.

1999-4 Christian M. Dahl and Svend Hylleberg: Specifying Nonlinear Econometric Models by Flexible Regression Models and Relative Forecast Performance.

CENTRE FOR NON-LINEAR MODELLING IN ECONOMICS

DEPARTMENT OF ECONOMICS - UNIVERSITY OF AARHUS - DK - 8000 AARHUS C - DENMARK
☎ +45 89 42 11 33 - TELEFAX +45 86 13 63 34

Working papers, issued by the Centre for Non-linear Modelling in Economics:

1996-3 Clive W.J. Granger and Niels Haldrup: Separation in Cointegrated Systems, Long Memory Components and Common Stochastic Trends.

1996-4 Morten O. Ravn and Martin Sola: A Reconsideration of the Empirical Evidence on the Asymmetric Effects of Money-Supply shocks: Positive vs. Negative or Big vs. Small?

1996-13 Robert F. Engle and Svend Hylleberg: Common Seasonal Features: Global Unemployment.

1996-14 Svend Hylleberg and Adrian R. Pagan: Seasonal Integration and the Evolving Seasonals Model.

1997-1 Tom Engsted, Jesus Gonzalo and Niels Haldrup: Testing for Multi-cointegration.

1997-7 Luca Fanelli: Estimating Multi-Equational LQAC Models with I(1) Variables: a VAR Approach.

1997-12 Niels Haldrup: A Review of the Econometric Analysis of I(2) Variables.

1997-14 Viggo Høst: Better Confidence Intervals for the Popoulation Mean by Using Trimmed Means and the Iterated Bootstrap?

1997-17 N.E. Savin and Allan H. Würtz: The Effect of Nuisance Parameters on Size and Power; LM Tests in Logit Models.

1997-18 Tom Engsted and Niels Haldrup: Multicointegration in Stock-Flow Models.

1998-6 Svend Hylleberg and Rikke Willemoes Jørgensen: A Note on the Estimation of Markup Pricing in Manufacturing.

1998-15 Henning Bunzel, Bent Jesper Christensen, Niels Haldrup, Svend Hylleberg, Viggo Høst, Peter Jensen og Allan Würtz: Udviklingslinier i Økonometrien.

1999-3 Niels Haldrup and Michael Jansson: Spurious Regression, Cointegration, and Near Cointegration: A Unifying Approach.

1999-4 Christian M. Dahl and Svend Hylleberg: Specifying Nonlinear Econometric Models by Flexible Regression Models and Relative Forecast Performance.